

データサイエンス概論I & II データサイエンス総論I & II

平均・分散・相関

九州大学 数理・データサイエンス教育研究センター

データの平均

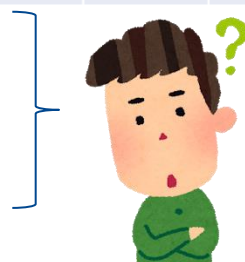
データの「代表値」の一つ

数多くのデータを「たった一つ」で代表させる

- Q: 3年10組の体重データを一言でいうとどんな感じ？

| | | | | | |
|------|------|------|------|------|------|
| 49.5 | 63.8 | 56.4 | 64.7 | 44.9 | 40.1 |
| 46.6 | 50.8 | 52.1 | 56.3 | 41.8 | 55.6 |
| 56.9 | 40.6 | 57.4 | 54.8 | 53.2 | 59.4 |
| 47.8 | 43.4 | 37.5 | 44.4 | 49.7 | 44.2 |
| 51.2 | 52.6 | 32.5 | 37.0 | 46.9 | 50.4 |

- A1: 最も重たいのは64.7kg
- A2: 最も軽いのは 32.5kg



そんな極端なケースを
言われてもなあ..

- A3: 平均すると約49.4kg

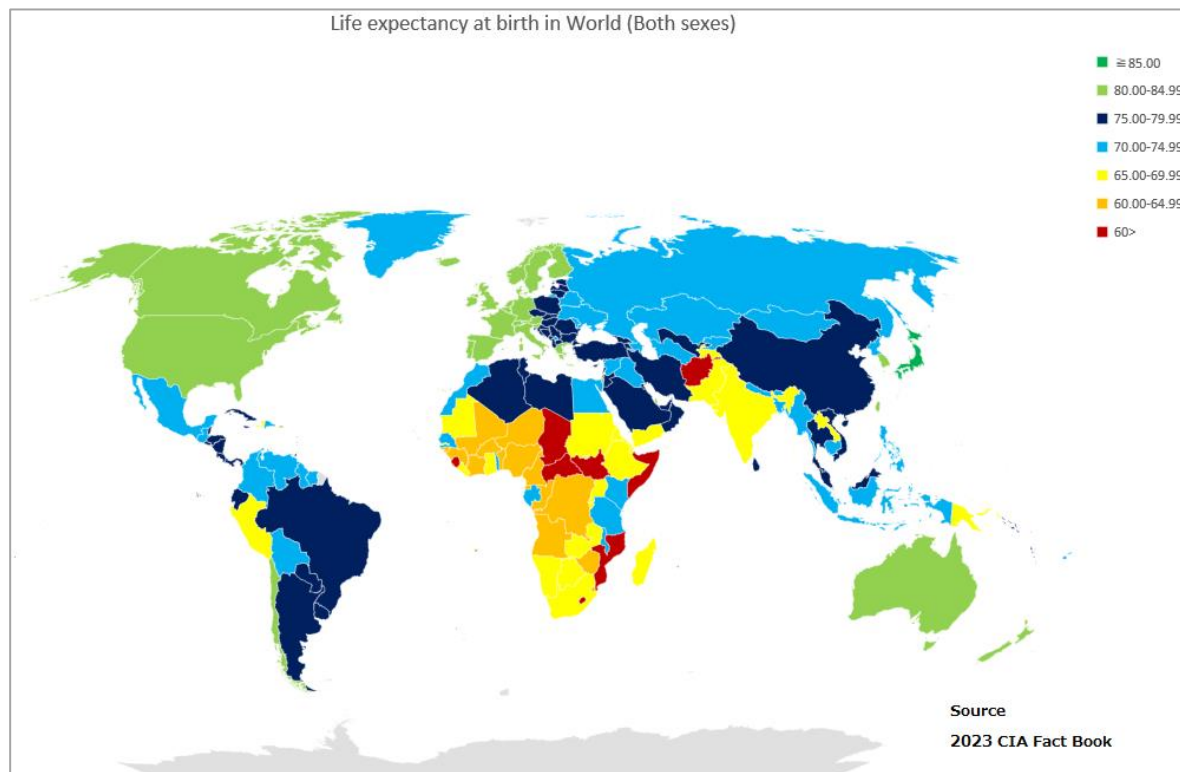


ああ、それぐらいの
体重の人がよくいるのね

平均は色々なところで使われる！

- 各国の平均寿命

- もちろん各国にはもっと短命・長命の人もあるが、傾向はわかる



<https://ja.wikipedia.org/wiki/平均寿命#/media/ファイル:2023年の国・地域別平均寿命（CIAファクトブックより）.png>

- 他にも色々

- 平均年収
- 英語テストの平均点
- 平均身長

データ集合に対する平均の計算法

- N 個のデータがあれば、基本は「全データを合計して」「 N で割る」
 - 正式には「算術平均」とか「相加平均」という名前がついている
- 例： $N = 5$ 人の体重 $\{62, 50, 49, 53, 73\}$ の場合
 - 平均 $= (62 + 50 + 49 + 53 + 73) / 5$

「ベクトル」データの集合の平均

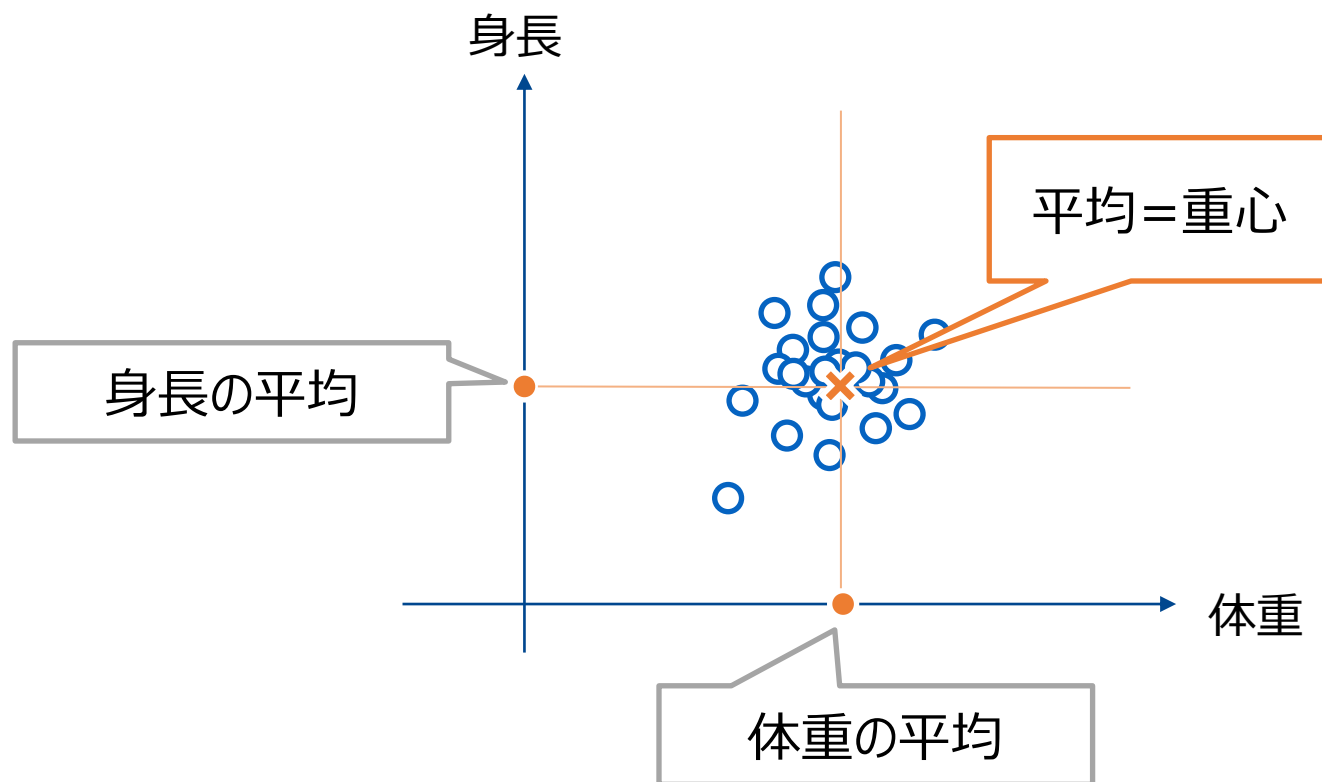
- 要素ごとに足して、足した個数(=データ数 N)で割るだけ

$$\begin{pmatrix} 62 \\ 173 \end{pmatrix} \quad \begin{pmatrix} 57 \\ 164 \end{pmatrix} \quad \begin{pmatrix} 65 \\ 171 \end{pmatrix} \quad \begin{pmatrix} 75 \\ 164 \end{pmatrix}$$

それぞれ合計して個数 N で割るだけ

- 何次元でも同じ
- 例： $N = 5$ 人の「(体重, 身長)の組」の場合
 - 平均 $= \left[\begin{pmatrix} 62 \\ 173 \end{pmatrix} + \begin{pmatrix} 50 \\ 162 \end{pmatrix} + \begin{pmatrix} 49 \\ 158 \end{pmatrix} + \begin{pmatrix} 53 \\ 156 \end{pmatrix} + \begin{pmatrix} 73 \\ 176 \end{pmatrix} \right] / 5$
 $= \begin{pmatrix} (62 + 50 + 49 + 53 + 73)/5 \\ (173 + 162 + 158 + 156 + 176)/5 \end{pmatrix}$

要するに各軸ごとに平均をとっているだけ



算術平均(要は「普通の平均」) を式で書くと...

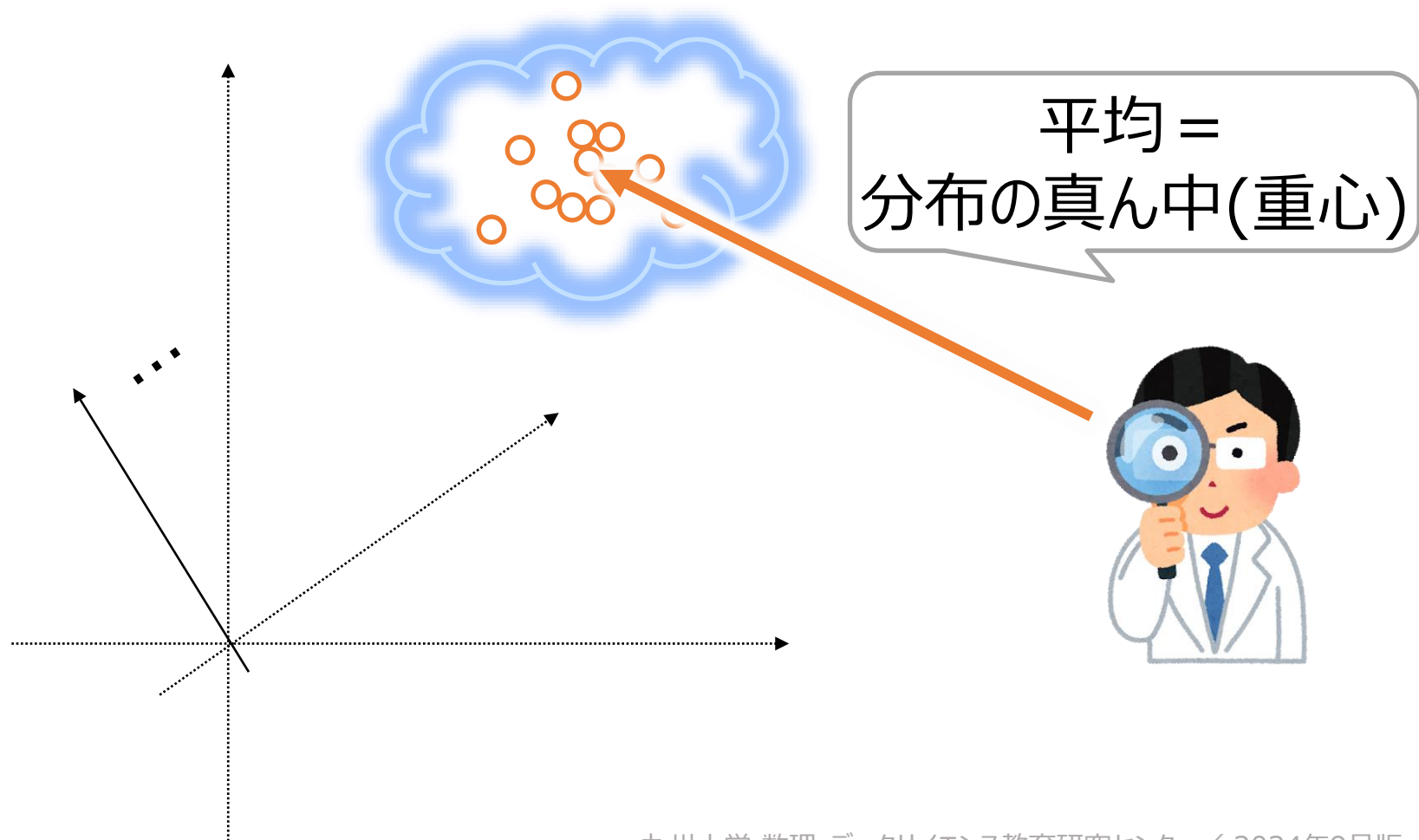
$$\bar{x} = \frac{1}{N}(x_1 + x_2 + x_3 + \cdots + x_N)$$

何度も+を
書くのが面倒

総和記号 Σ を使うと短く書ける

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$$

平均は分布の形を探る「第一歩」： データは平均を中心として広がっている！







ちょっと凝った平均： 加重平均 (1/3)

- 重みを付けて算術平均

時々軽めに申告するので
あてにならない…

重みの例

確からしさ w_i

| | | | |
|---|--|---|---|
|  |  |  |  |
| $x_1, x_2, x_3, \dots, x_N$ | | | |
| ↑ | ↑ | ↑ | ↑ |
| $\begin{pmatrix} 62 \\ 173 \end{pmatrix}$ | $\begin{pmatrix} 57 \\ 164 \end{pmatrix}$ | $\begin{pmatrix} 65 \\ 171 \end{pmatrix}$ | $\begin{pmatrix} 75 \\ 164 \end{pmatrix}$ |
| 1 | 0.9 | 0.95 | 0.1 |





$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \dots + w_N x_N}{w_1 + w_2 + w_3 + \dots + w_N} = \frac{\sum w_i x_i}{\sum w_i}$$

ちょっと凝った平均： 加重平均 (2/3)

- 算術平均は加重平均の特殊な場合

重みの例

確からしさ w_i





| | | | |
|---|--|---|---|
|  |  |  |  |
| x_1 | x_2 | x_3 | x_N |
| ↑ | ↑ | ↑ | ↑ |
| $\begin{pmatrix} 62 \\ 173 \end{pmatrix}$ | $\begin{pmatrix} 57 \\ 164 \end{pmatrix}$ | $\begin{pmatrix} 65 \\ 171 \end{pmatrix}$ | $\begin{pmatrix} 75 \\ 164 \end{pmatrix}$ |
| 1 | 1 | 1 | 1 |

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_N x_N}{\underbrace{w_1 + w_2 + w_3 + \cdots + w_N}_{1\text{が}N\text{個}}} = \frac{1}{N} \sum x_i$$

ちょっと凝った平均： 加重平均 (3/3)

- 全員「あてにならない」場合も、普通の加重平均に

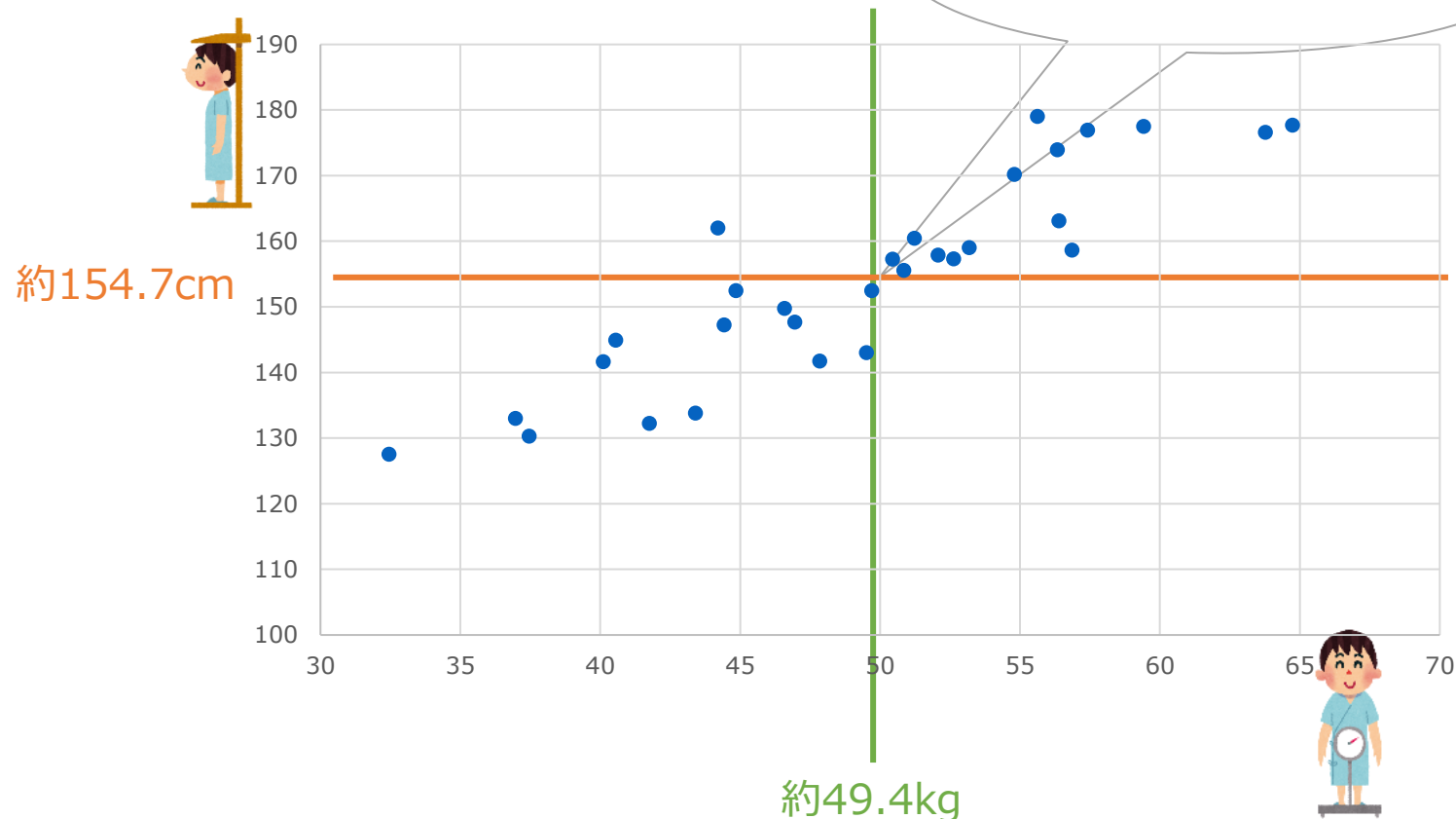
重みの例

| | | | |
|---|--|---|---|
|  |  |  |  |
| x_1 | x_2 | x_3 | x_N |
| ↑ | ↑ | ↑ | ↑ |
| $\begin{pmatrix} 62 \\ 173 \end{pmatrix}$ | $\begin{pmatrix} 57 \\ 164 \end{pmatrix}$ | $\begin{pmatrix} 65 \\ 171 \end{pmatrix}$ | $\begin{pmatrix} 75 \\ 164 \end{pmatrix}$ |
| 確からしさ w_i | 0.1 | 0.1 | 0.1 |

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2 + w_3 x_3 + \cdots + w_N x_N}{\underbrace{w_1 + w_2 + w_3 + \cdots + w_N}_{0.1 \text{ が } N \text{ 個}}} = \frac{1}{N} \sum x_i$$

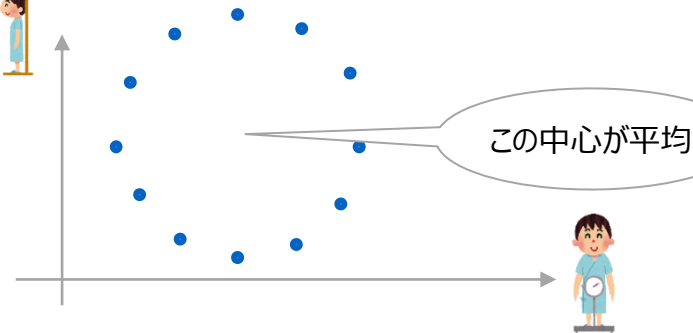
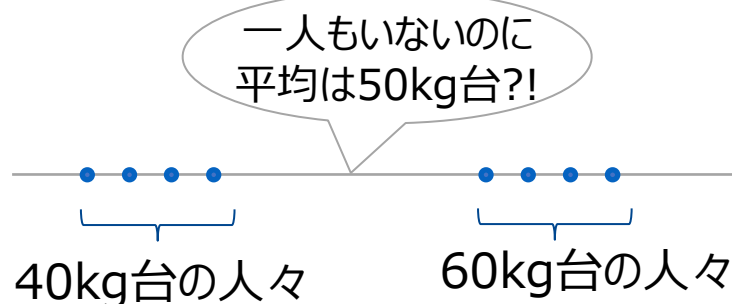
平均は代表値としてふさわしくない場合がある(1/3)

平均と全く同じデータがあるわけではない



平均は代表値としてふさわしくない場合がある(2/3) 「よくある値」ですらないことも

!



なので、この話は
いつもそうとは限らない

- A3: 平均すると約49.4kg



ああ、それぐらいの
体重の人がよくいるのね

平均は代表値としてふさわしくない場合がある(3/3)

はずれ値に弱い

- **はずれ値** = 例外的なデータ(次スライド)



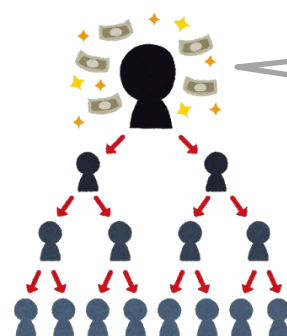
1人は巨大ロボだった...

- 例 : $N = 5$ 人の体重{62, 50, 49, 53, 550000}
- 平均でおよそ110043kg (!?)

- 「平均」を悪用



平均1千万
儲かるよ



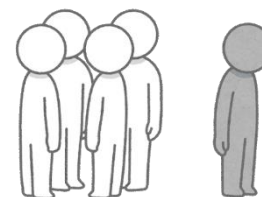
+100億円

-100万円

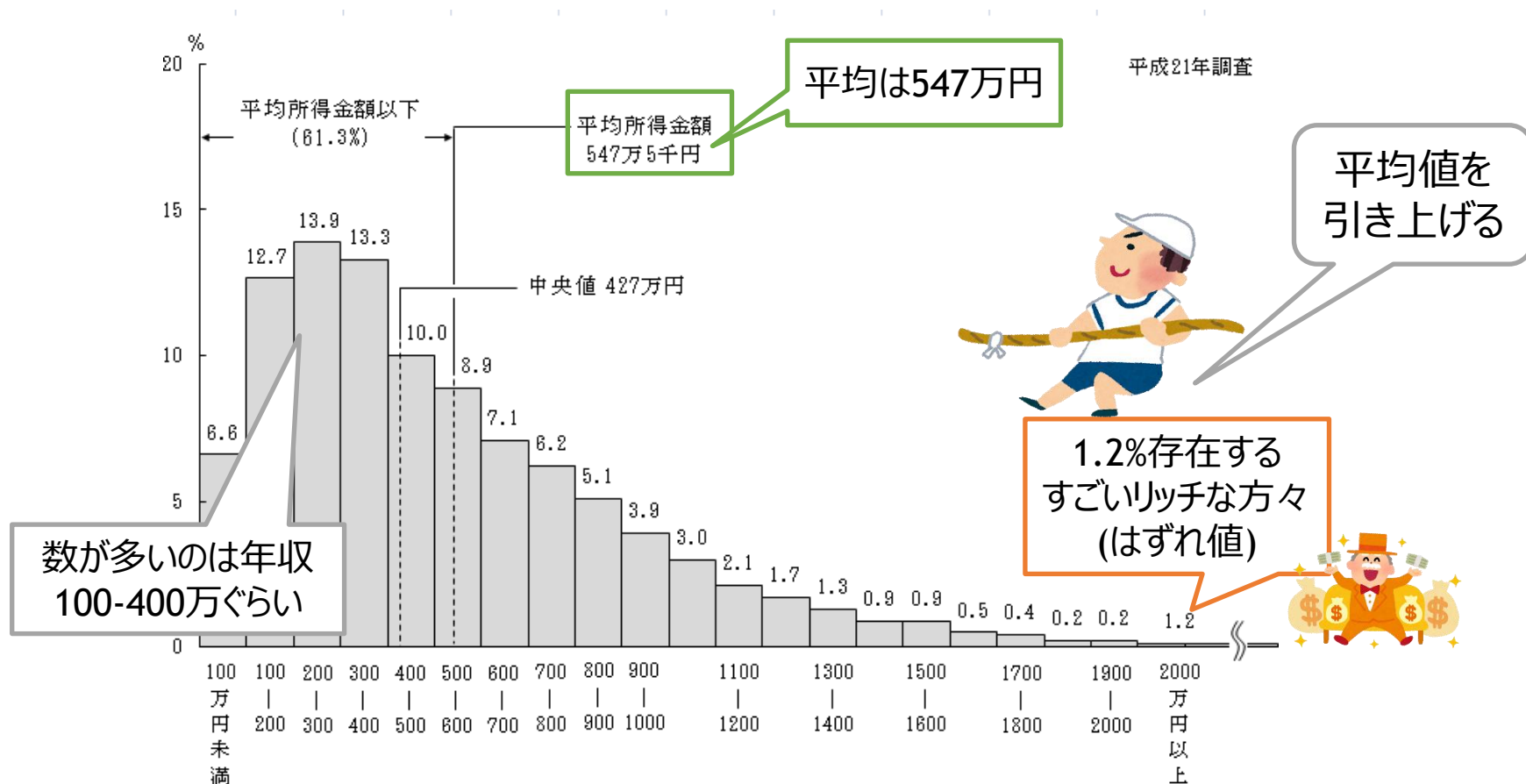
- ウソではないかもしれないが、ごく少数の人だけが莫大な利益を上げ、残り大多数は大損している可能性も

はずれ値(outlier) = 例外的な値

- 一般的なデータと著しく異なる値を持つデータ
 - 結構よくある
- はずれ値は色々な原因で発生する
 - 測定ミス
 - 測定機器の故障
 - イイカゲンな回答者
 - 異常現象
 - 希少現象(めったに起きない現象)
 - 想定外の現象・初めて発生した現象



はずれ値の影響： 日本人の年収(ヒストグラム)を例に



平均以外の「代表値」

平均だけじゃない

平均以外の代表値(1/2) : 中央値 (メディアン)

- 数値の大きさの順に並べた時に, 真ん中に来る値
 - 順位データ (ex. アンケート結果) にも使える
 - ベクトルデータには普通使わない

複数のベクトルを大きい順に並べる一般的な方法がないから
(どうしても使いたければ, 次元毎に独立にメディアンを求める?)

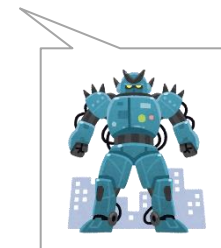
- 例 : $N = 5$ 人の体重 $\{62, 50, 49, 53, 73\}$ の場合
 - 並べ替えると, 49, 50, 53, 62, 73
 - なのでメディアン $\tilde{x} = 53$

N が偶数の場合はちょっと工夫が必要.
ex. 6 人の場合は, 3 位と 4 位の平均

- メディアンのよいところ... 極端な「はずれ値」の影響を受けない
 - $\{62, 50, 49, 53, 185\}$ となっても, $\tilde{x} = 53$
 - $\{62, 50, -2, 53, 73\}$ となっても, $\tilde{x} = 53$
- 困ったところ (?)
 - $\{1, 2, 3, 200, 201\}$ ならば, $\tilde{x} = 3$

中央値（メディアン）は「はずれ値」に強い

- 例： $N = 5$ 人の体重{62, 50, 49, 53, 550000}の場合
 - 並べ替えると, 49, 50, 53, 62, 550000
 - なので, 中央値は53
- はずれ値に全く影響されていない!

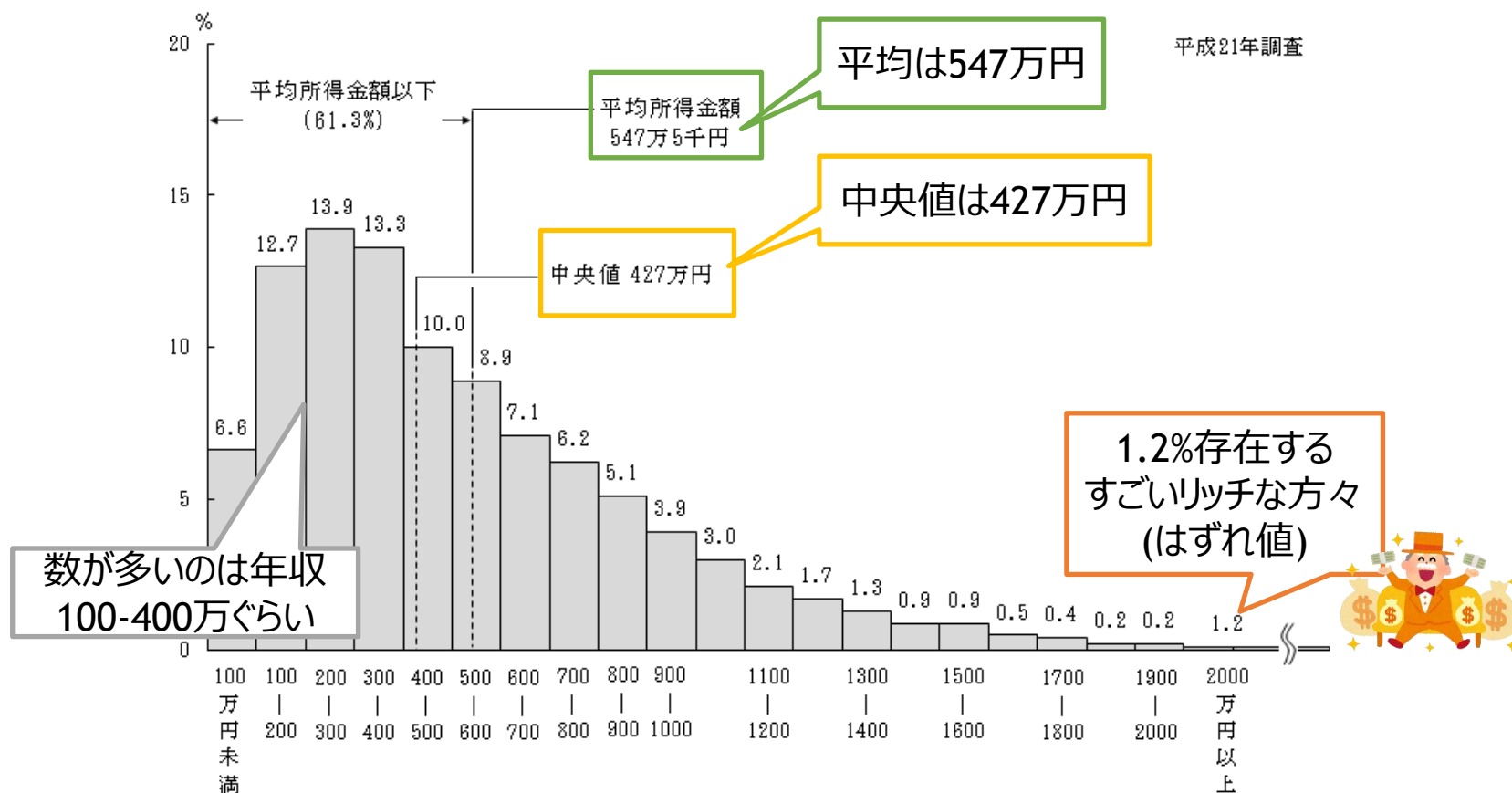


並みはずれた体重



中央値（メディアン）は「はずれ値」に強い： 日本人の年収（ヒストグラム）を例に

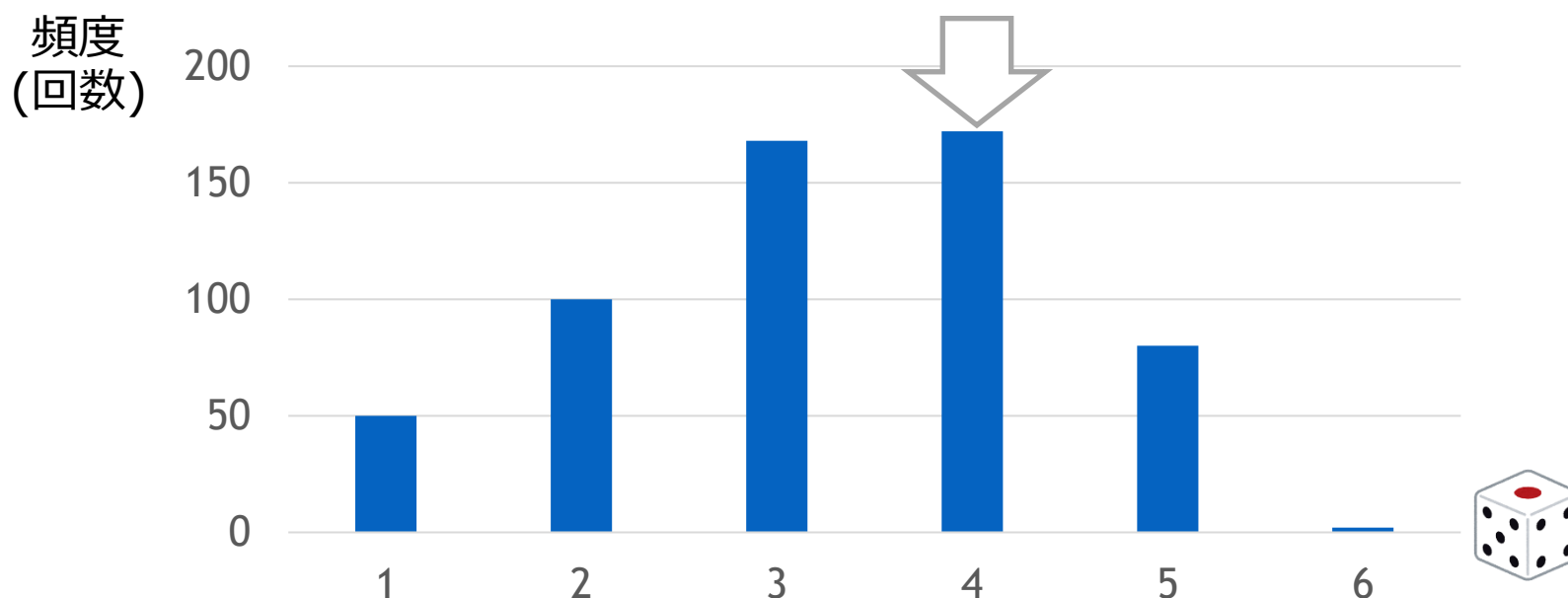
- 中央値のほうがはずれ値に影響されにくそう



厚生労働省 2009年調査 <https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa09/2-2.html>

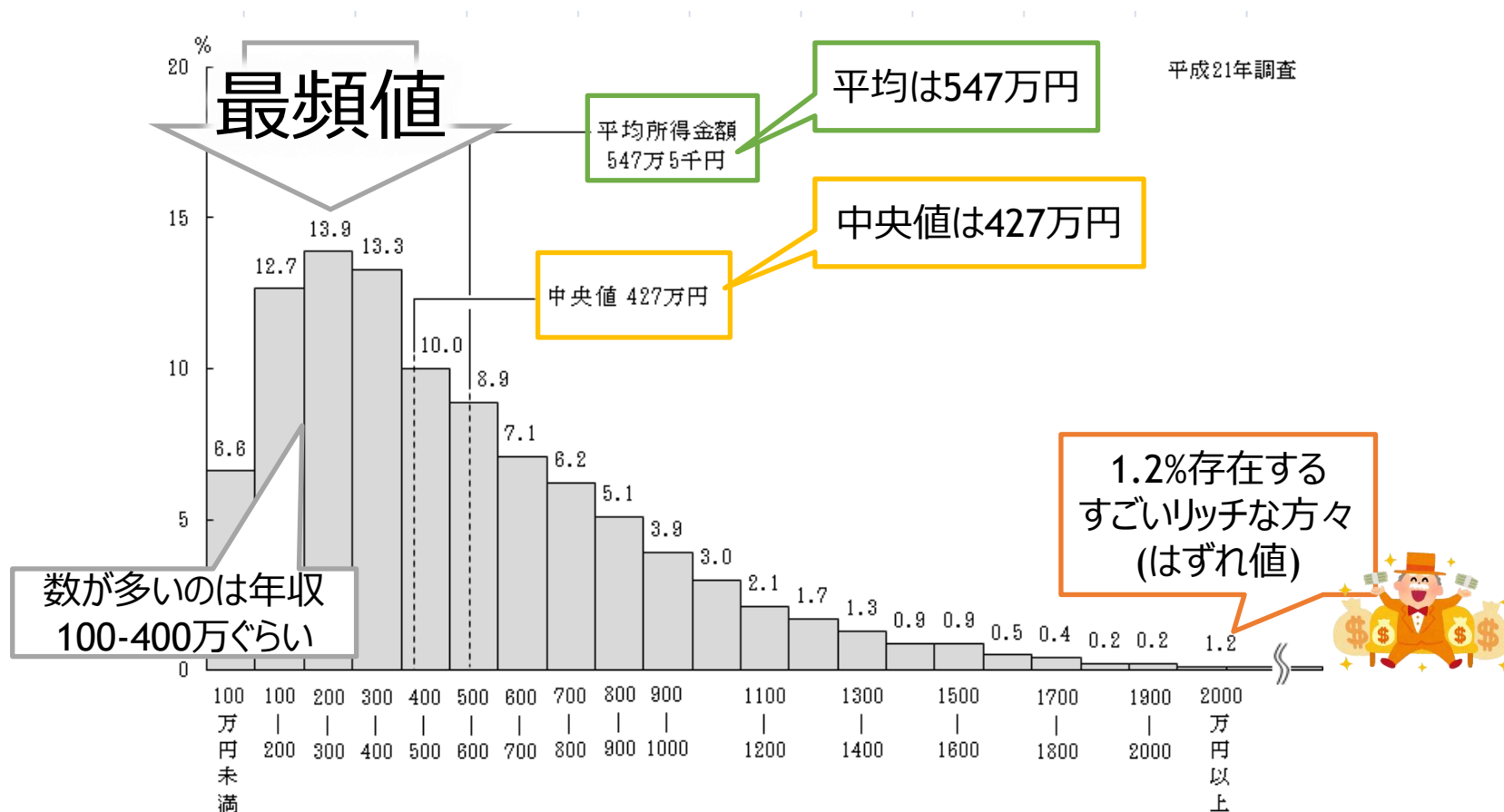
平均以外の代表値(2/2)： 最頻値 (モード)

- 最も頻度が高い=最も出やすいデータ=ヒストグラムのピーク
 - ズルいサイコロの最頻値は“4”



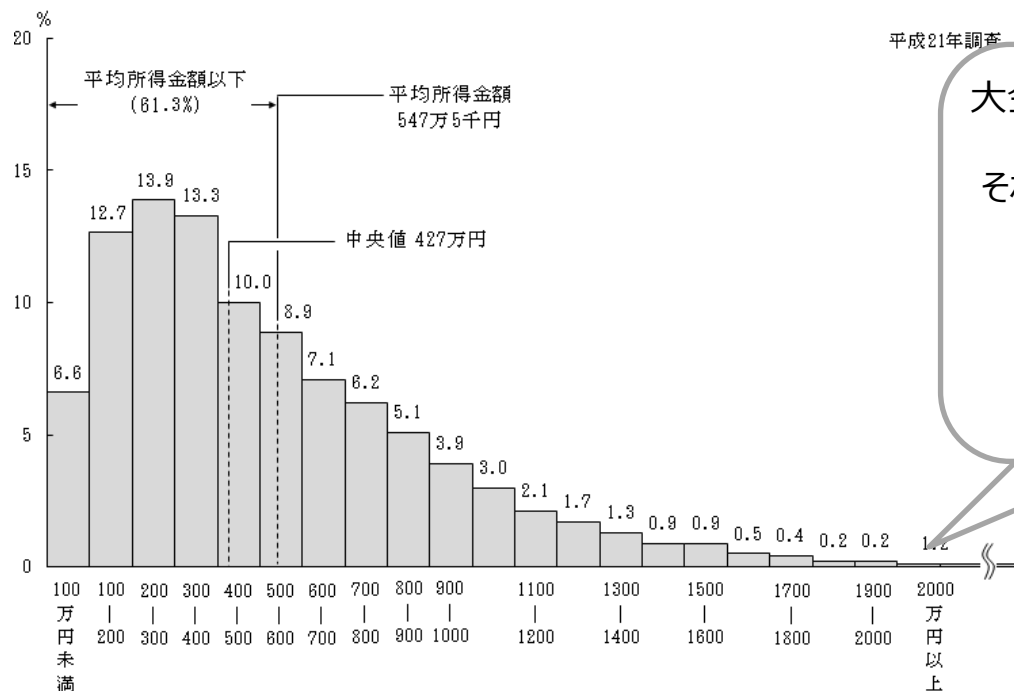
- カテゴリデータ(ex. バスの番号)にも使える！

最頻値（モード）も「はずれ値」に強い： 日本人の年収（ヒストグラム）を例に



確かに、中央値や最頻値は「はずれ値」に強い： しかし本当にそれでいいのか？(1/2)

- はずれ値でも「本当のデータ」の場合もある



大金持ちは日本にも実在!

それなのに日本の代表値
計算から外して、
無視してよいのか？



厚生労働省 2009年調査 <https://www.mhlw.go.jp/toukei/saikin/hw/k-tyosa/k-tyosa09/2-2.html>

確かに、中央値や最頻値は「はずれ値」に強い： しかし本当にそれでいいのか？(2/2)

- 「はずれ値」として見捨てていいかは場合による
 - ex. 毎月のインフルエンザ死亡者数（10万人あたり）

| 1月 | 2月 | 3月 | 4月 | 5月 | 6月 | 7月 | 8月 | 9月 | 10月 | 11月 | 12月 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 2.1 | 7.3 | 9.1 | 4.6 | 1.6 | 0.5 | 0.1 | 0.1 | 0.1 | 0.1 | 0.3 | 1.4 |

小→大の順に並べ替え

| 7月 | 8月 | 9月 | 10月 | 11月 | 6月 | 12月 | 5月 | 1月 | 4月 | 2月 | 3月 |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0.1 | 0.1 | 0.1 | 0.1 | 0.3 | 0.5 | 1.4 | 1.6 | 2.1 | 4.6 | 7.3 | 9.1 |

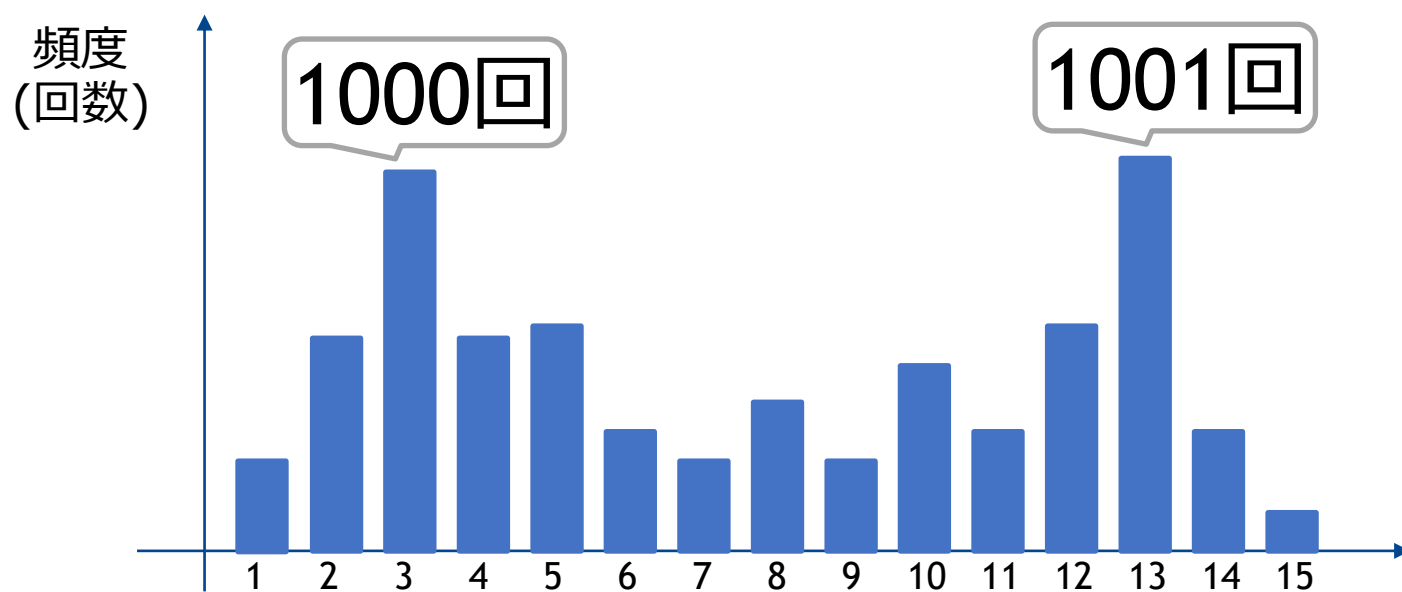
中央値 0.5 or 1.4

※データが偶数個なので2つある

はずれ値として
中央値に影響なし
→いいのか？

最頻値（モード）には別の弱点も

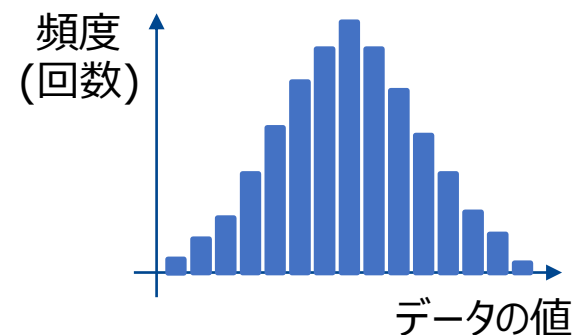
- 似たピークが複数ある場合，わずかな差異で全く異なる最頻値に



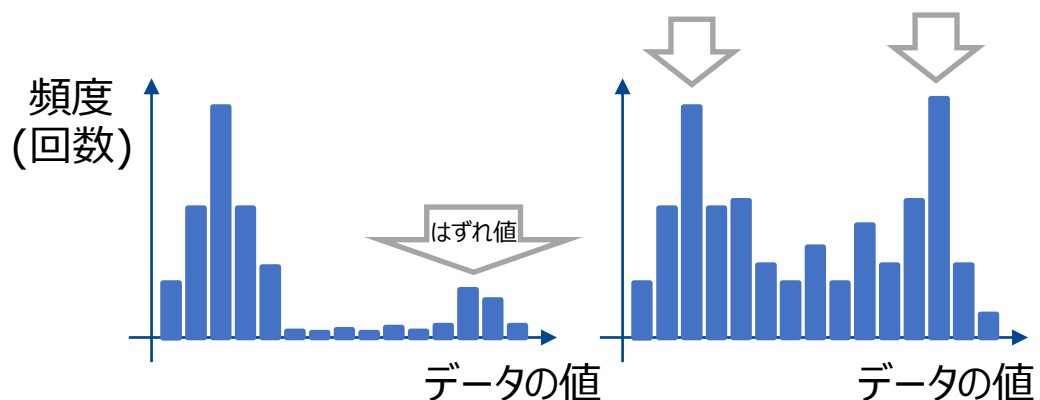
- 最頻値は13だけど，頻度のわずかな変化で3にもなりうる

代表値の使い方に関する「お勧め」： 全ての代表値(平均/中央値/最頻値)を出してみる

- もし、全部がだいたい同じなら
 - はずれ値があまりない
 - ヒストグラムで書くと、左右対称な山とかになってそう→



- 値が結構違うなら
 - はずれ値がある
 - ヒストグラムで書くと山が複数
 - 山が「へ」の字状に偏っている
 - ...



- 以上に加え、分散(後述)を出してみるのも効果的

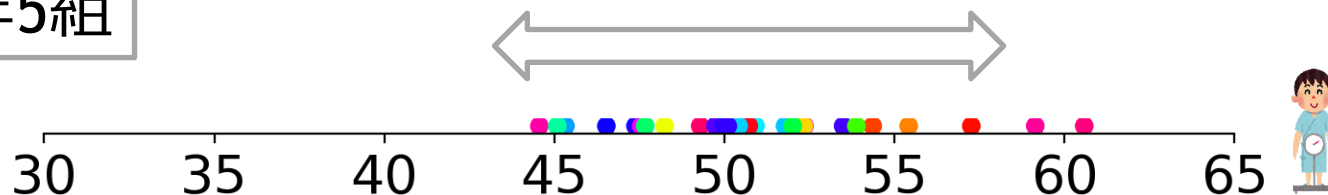
データの分散

分布(=データ集合)の性質を記述する第二歩.

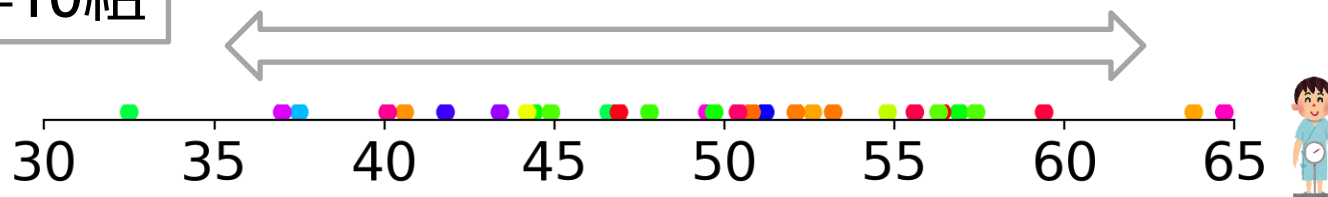
データのばらつき： 体重データの場合

- ばらつき = データの**広がり**具合 = データの**変動**具合

3年5組



3年10組



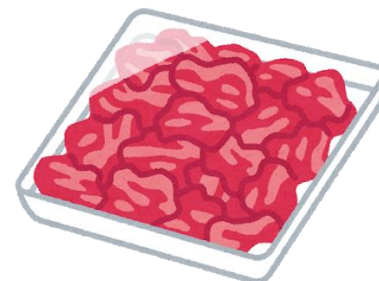
- 3年10組のほうが体重が「ばらついて」いる！
 - 平均体重はどちらも同じぐらい(約50kg)

全国物価統計調査 平成9年全国物価統計調査 大規模店舗編

● 福岡県の牛肉価格だけ取り出してみた

| | 地域 | 調査数 | 平均価格 | 標準偏差 |
|---------|---------|-----|-------|-------|
| 牛肉, 国産品 | 福岡県 町村 | 21 | 517.2 | 122.9 |
| | 福岡県 福岡 | 211 | 530.3 | 129.2 |
| | 福岡県 北九州 | 139 | 507.7 | 148.8 |
| | 福岡県 筑豊 | 35 | 514.9 | 181 |
| | 福岡県 筑後 | 52 | 488.3 | 143.2 |
| 牛肉, 輸入品 | 福岡県 町村 | 19 | 268.1 | 96.9 |
| | 福岡県 福岡 | 145 | 304.1 | 106 |
| | 福岡県 北九州 | 123 | 266.8 | 112.5 |
| | 福岡県 筑豊 | 27 | 251.7 | 86.6 |
| | 福岡県 筑後 | 42 | 309.3 | 122.9 |

ばらつき
具合



筑豊のスーパー,
国産牛肉の価格の
店ごとの「ばらつき」が
他地域より大きい!
でも輸入牛肉については
他地域より小さい..
不思議...

↓全データはこちらから入手可能

<https://www.e-stat.go.jp/dbview?sid=0000100087>

↓最新の物価統計調査も手に入りますよ

<http://www.stat.go.jp/data/kouri/doukou/index.html>

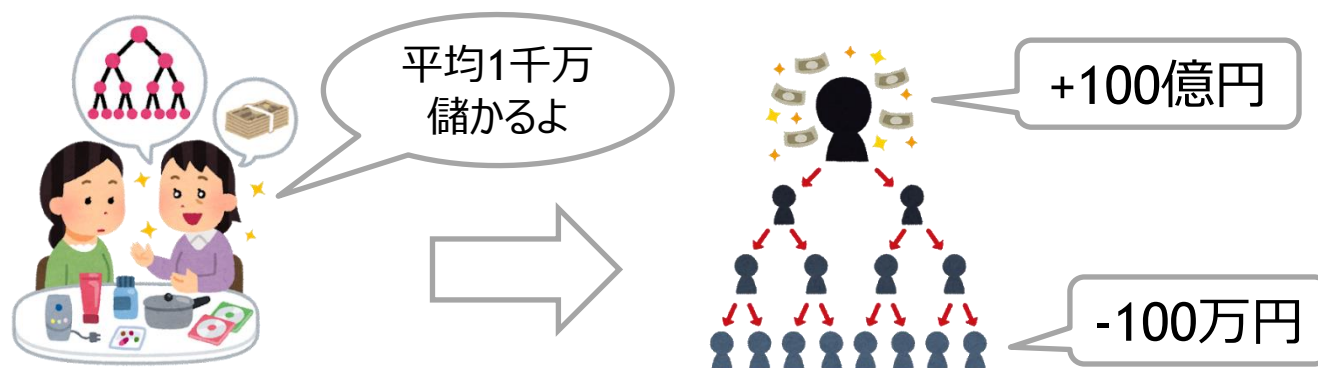
「ばらつき」を考える理由

- 代表値（平均・中央値・最頻値）に加えて、データの「ばらつき」を知りたいことは多い
 - 例1: アンケート「この曲好きですか？ 1(大嫌い), 2, 3, 4, 5(大好き)」
 - 「平均3で, ばらつき小」→ 大多数が3 → みんな普通 → あまり売れなそう
 - 「平均3で, 1から5までばらつく」→ 好みの別れる曲
 - ↑ 平均は同じでもばらつきが違えば, 状況は全然違う
 - 例2: あるダイエット食品で減った体重
 - 「平均ゼロ, ばらつき小」→ 大多数がゼロ → 多くの人には効果なし
 - 「平均ゼロ, -10kgから+10kgまでばらつく」→ 非常に効果ある人もいるが, 逆効果大の人もいる（効果に個人差がありすぎて危険）
 - 「平均 -5kg, ばらつき小」→ 大多数が -5kg → 多くの人に効果あり



「ばらつき」を考える理由： 要するに「代表値だけではわからないことが多い」

- 思い出そう：「平均」を悪用



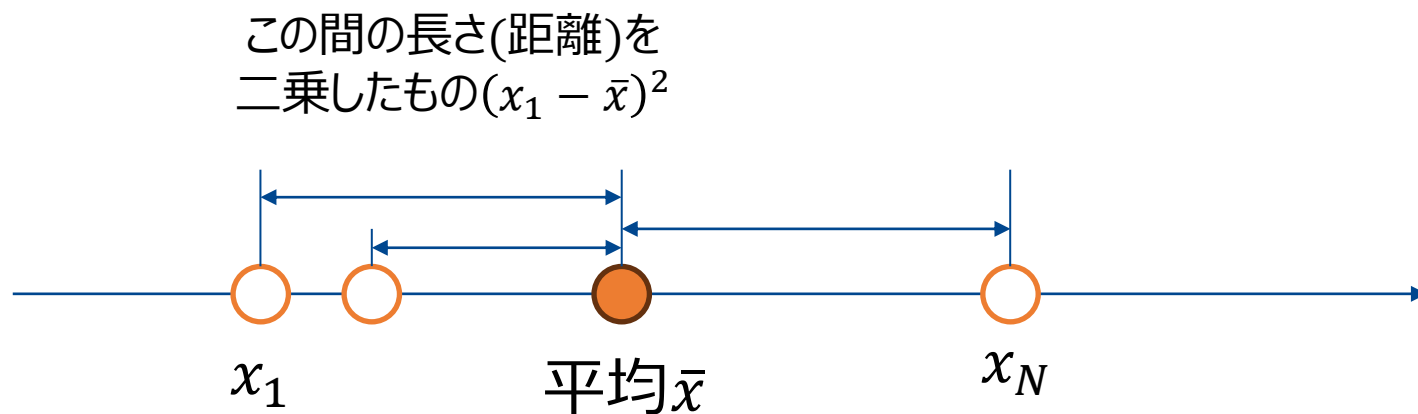
- ウソではないかもしれないが、ごく少数の人だけが莫大な利益を上げ、残り大多数は大損している可能性も

儲けの「ばらつき」を確認すれば「悪用」がわかる！
全員が1千万円儲けているわけではない！



分散 = 数の集合 (例えば体重の集合) のばらつきを測る方法

- 全データが平均的に「平均 \bar{x} とどれくらい離れているか？」ではどう？
 - ※ただし離れ具合は「二乗距離」で評価。また「距離」については後述



- 分散が大きい→平均値と大きく違う数が多い→広がっている

このアイデアを式で表すと…

- 数の集合 x_1, x_2, \dots, x_N の分散
 - = 「(算術)平均値との差の二乗」の平均

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + \dots + (x_N - \bar{x})^2}{N}$$

算術平均

$$= \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

なんで二乗？ 絶対値とかでもよさそうだし…

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_N - \bar{x})^2}{N}$$

その素晴らしい疑問に対する答えは、付録に

練習

- $1, 1, 1, 1, 1$ の分散は？
- $1, 5, 4, 2, 8$ の分散は？

分散, ちょっとした話(1/2)

- 全部の数が一様に Δ だけプラスされても, 分散は同じ
 - 値が x_i から $x_i + \Delta$ になったとすると, 平均は \bar{x} から $\bar{x} + \Delta$ になるので,

$$\frac{1}{N} \sum_{i=1}^N ((x_i + \Delta) - (\bar{x} + \Delta))^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \sigma^2$$



分散, ちょっとした話(2/2)

- では全部の数が一様に α 倍になったらどうなる？
 - 値が x_i から αx_i になったとすると, 算術平均は \bar{x} から $\alpha \bar{x}$ になるので,

$$\frac{1}{N} \sum_{i=1}^N (\alpha x_i - \alpha \bar{x})^2 = \frac{\alpha^2}{N} \sum_{i=1}^N (x_i - \bar{x})^2 = \alpha^2 \sigma^2$$

α^2 倍に!



練習

- 1, 1, 1, 1, 1 の分散は？

- 1, 5, 4, 2, 8 の分散は？

+1000

- 1001, 1005, 1004, 1002, 1008の分散は？

×10

- 10, 50, 40, 20, 80 の分散は？

データの標準偏差

標準偏差は分散よりもわかりやすい

標準偏差 = 分散の平方根

- 例
 - 分散が100なら, 標準偏差は10
 - 分散が2なら, 標準偏差は $\sqrt{2}$
 - 分散が0なら, 標準偏差も0
- なので, 標準偏差を二乗したら分散
 - 標準偏差が10なら, 分散は100
 - 標準偏差が $\sqrt{2}$ なら, 分散は2

標準偏差は何のためにある？

- 分散が大きい（小さい）ければ標準偏差も大きい（小さい）
 - よって、分散と同じように、全データの「ばらつき」を表す値
- 標準偏差は分散よりもわかりやすい
 - 分散を求める際、二乗して「ずれ」を求めている
 - 体重のデータや平均には50kgのように単位kgがつく
 - しかし分散を求めるときには二乗したので、 62.0kg^2 のような妙な単位に
 - 標準偏差は、その平方根なので、より「ずれ」をシンプルに表す
 - 単位は再びkgに戻る （ $62.0\text{kg}^2 \rightarrow \sqrt{62.0}\text{kg}$ ）
 - なので、 $50\text{kg} \pm \sqrt{62.0}\text{kg}$ のような感じで、ばらつきの範囲を使う際にも使える！

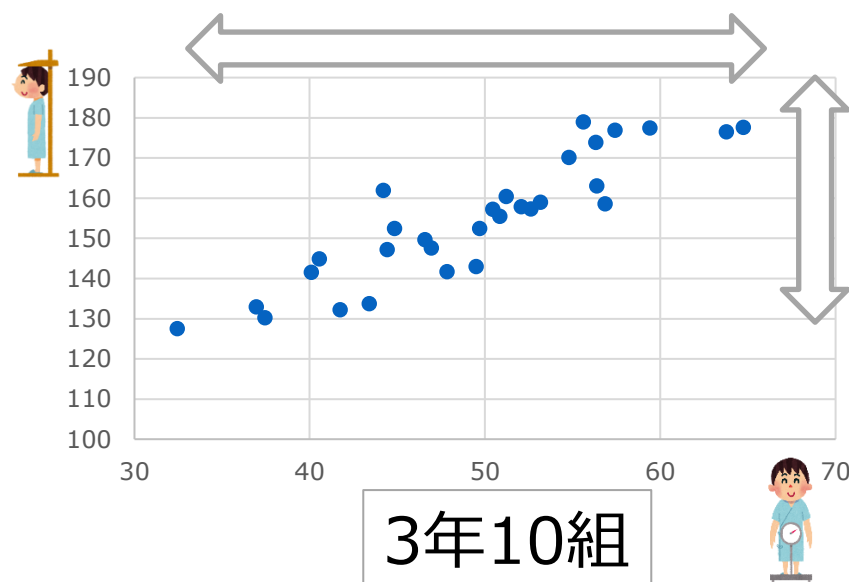
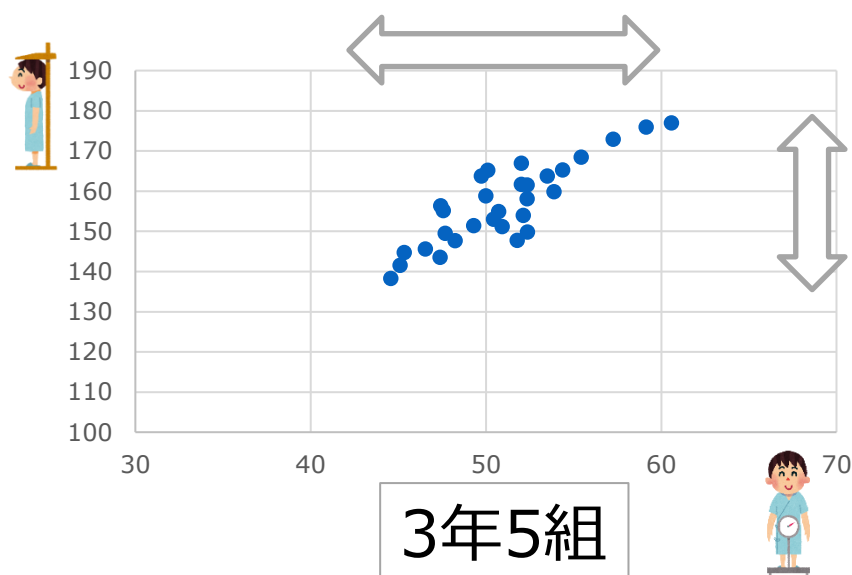
なお、標準偏差 ≠ 偏差値。偏差値については付録に

ベクトルデータの分散は？

多変量解析のステップ1

ベクトルデータの分散は？(1/3)

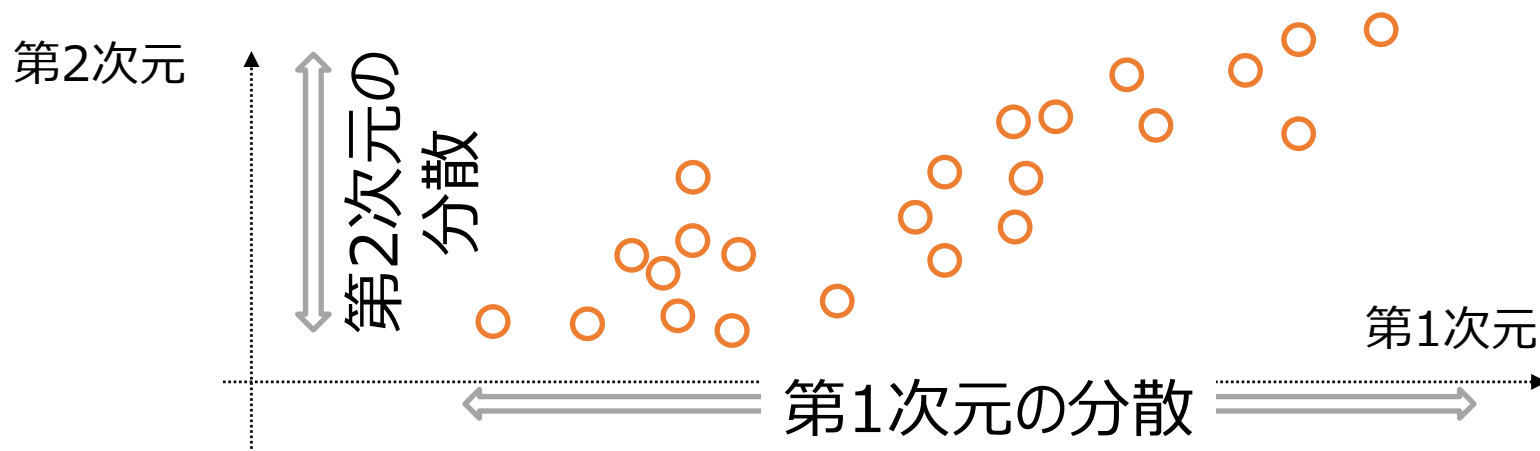
まずは各軸ごとの分散を考える



- 3年10組の方が，身長も体重もばらつきが大きい…
 - 平均体重・平均身長はどちらも同じぐらい(約50kg, 155cm)

ベクトルデータの分散は？(2/3)

まずは各軸ごとの分散を考える

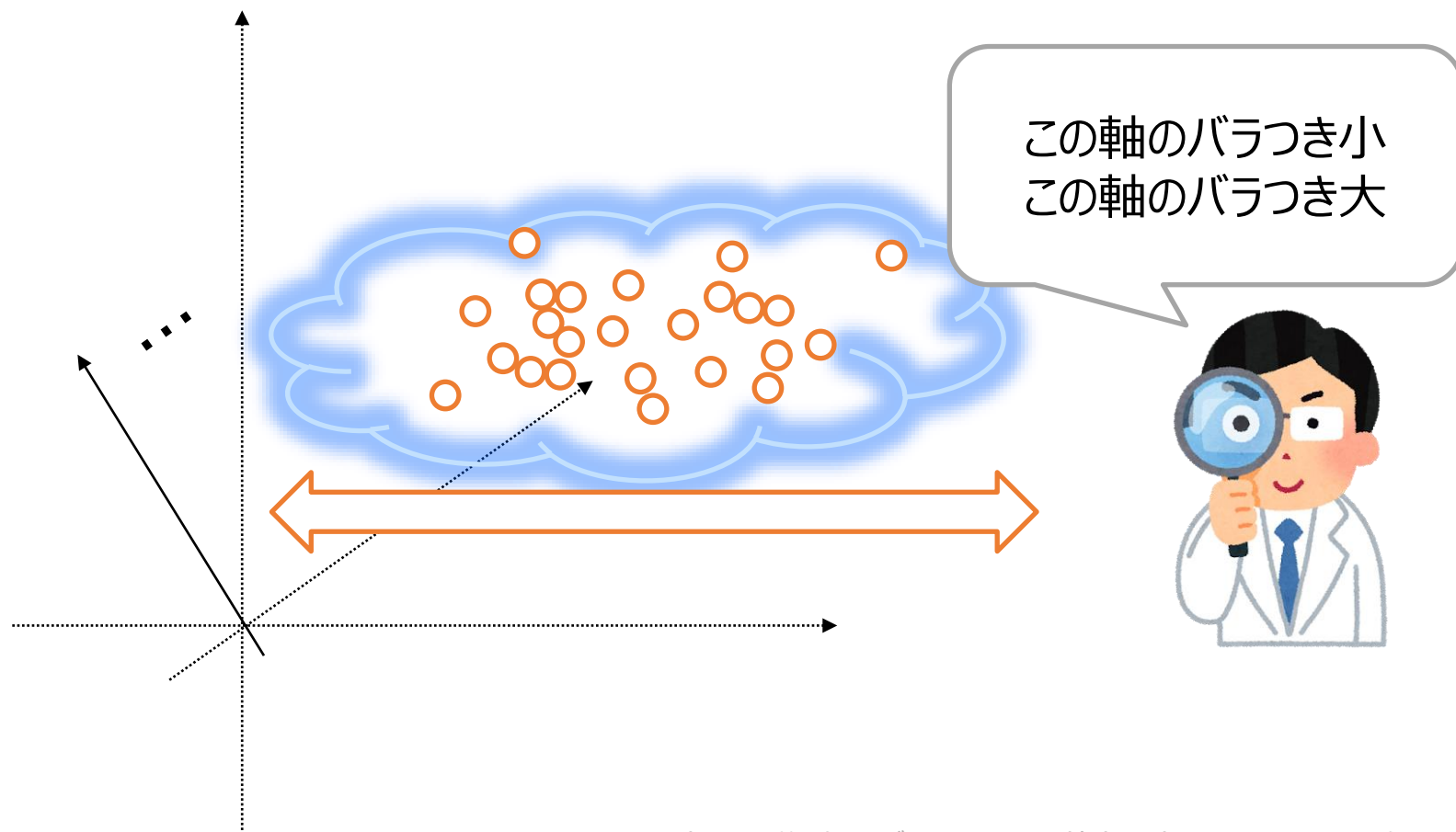


- 上図では、「第1次元の分散 > 第2次元の分散」

ベクトルデータの分散は？ (3/3)

まずは各軸ごとの分散を考える

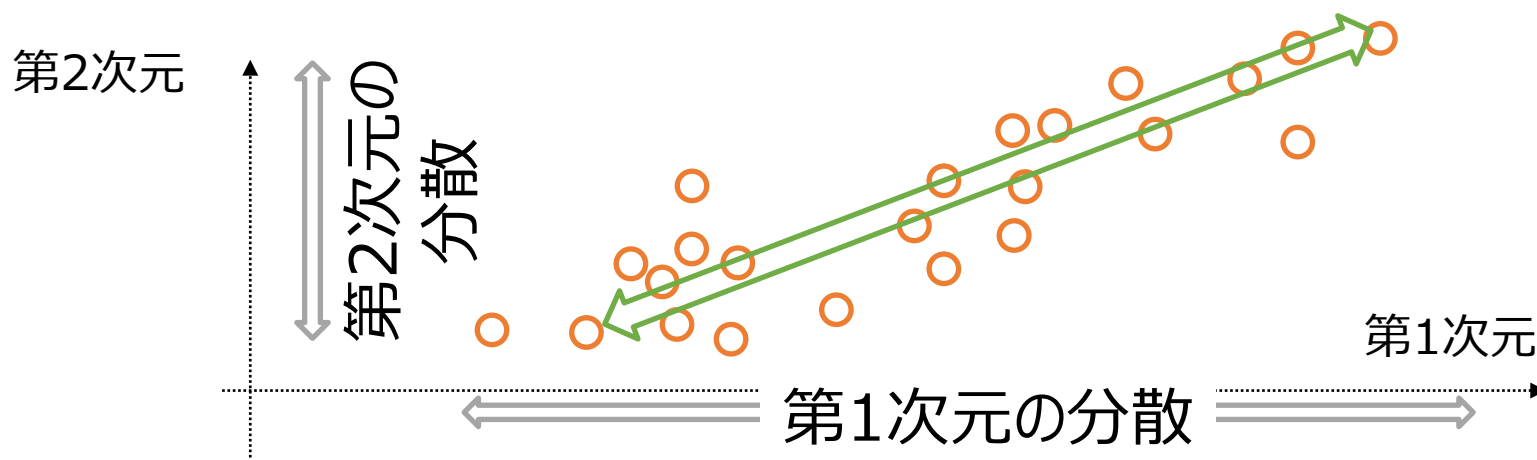
- このように各軸での分散を見ることで，各軸でデータがどれくらい広がっているかがわかる



「軸ごとの分散でいいのか？」と 疑問に思った人は大正解



- **斜め方向**（例えば、分布が最も広がっている方向）の分散を見るべき場合もあります！



- そこで次は「相関」について考えましょう

相関

「身長が高ければ、体重も重い」傾向

データの広がり方(=分散)に潜む関係～相関

Case 1

身長と体重は？
身長と数学の点数は？
身長と400m走のタイムは？

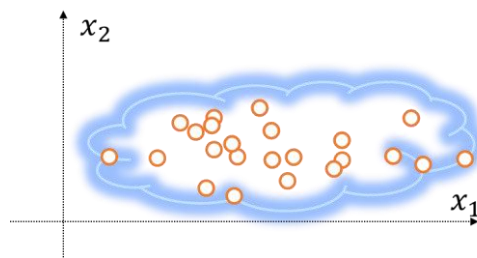
Case 2

Case 3

データの広がり方(=分散)に潜む関係～相関

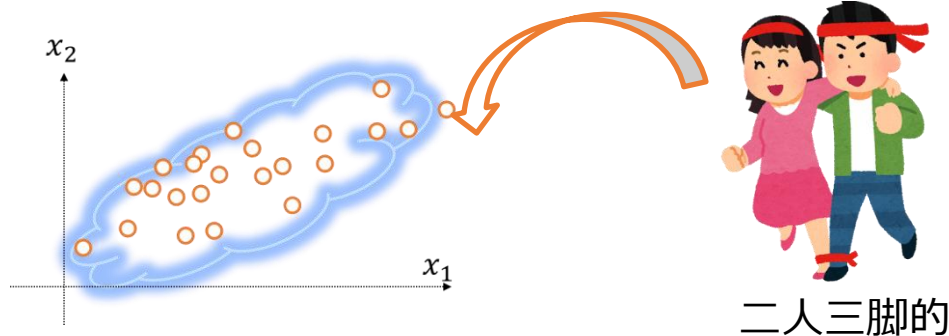
- Case 1: 無相関

- $x_1 \rightarrow$ 大, $x_2 \rightarrow$ 特段の傾向無し
- 要するに, x_1 と x_2 は無関係
- 身長と数学の点数



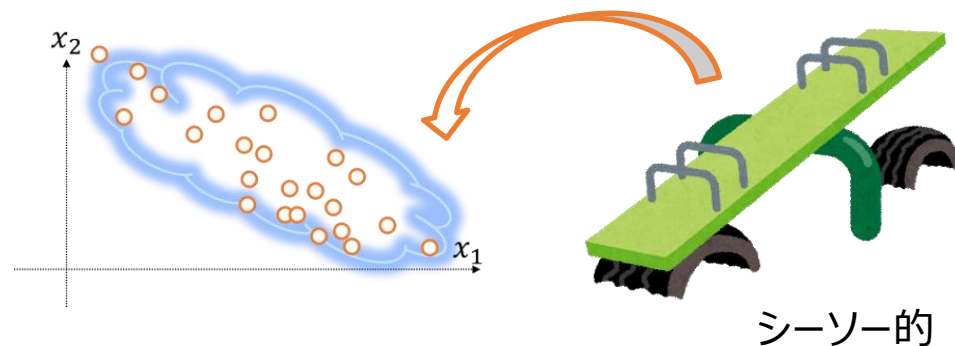
- Case 2: 正の相関

- $x_1 \rightarrow$ 大, $x_2 \rightarrow$ 大
- 身長と体重



- Case 3: 負の相関

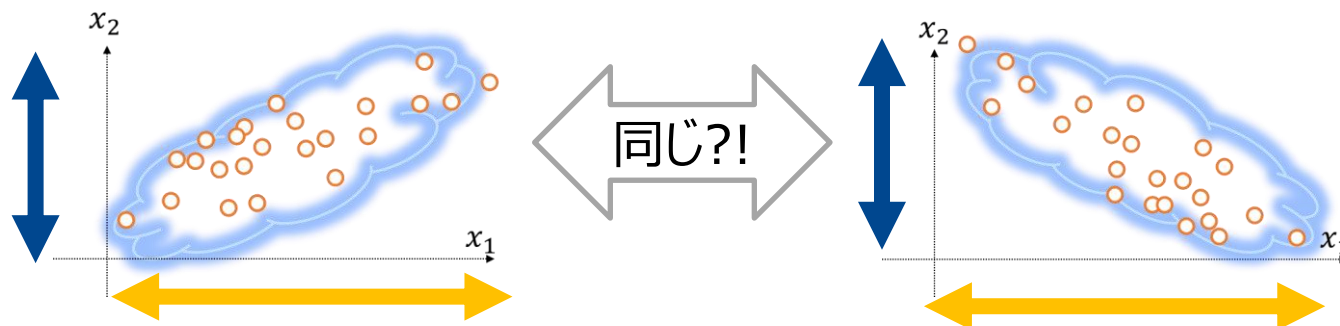
- $x_1 \rightarrow$ 大, $x_2 \rightarrow$ 小
- 身長と400m走のタイム



「収入 vs エンゲル係数」も

相関とは？

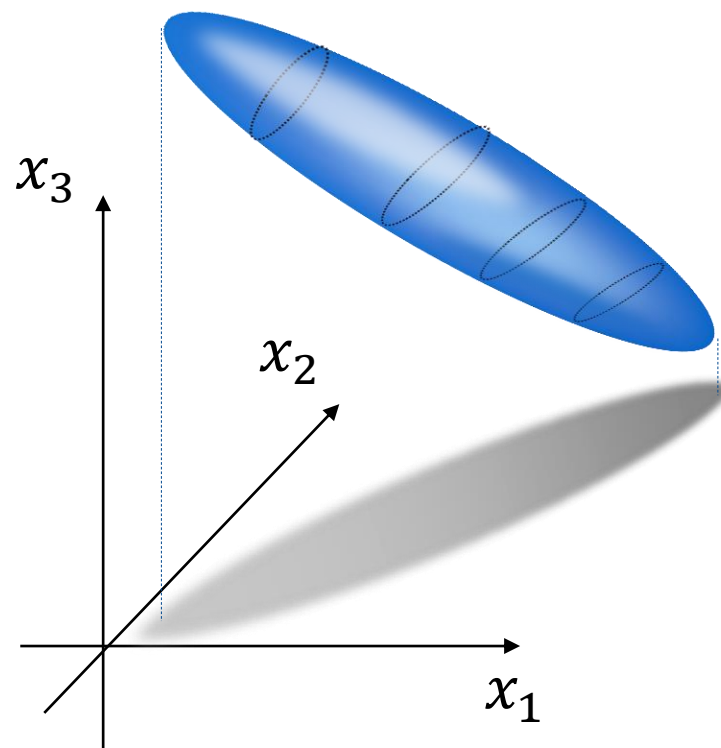
- データの要素間の関係・傾向
 - Ex. $x_1 \rightarrow$ 大なら $x_2 \rightarrow$ 大 , とか
 - これは平均では記述できない
 - 各軸独立の分散では記述できない
 - 各軸独立でみると...



- 主成分分析_(後述)や回帰分析_(後述)もある意味で相関を見つけてくれる

多次元($d > 2$)ベクトルの相関

- もちろん同じようなことがわかる
- 関係はより複雑になりうる
 - 右図では
 $x_1 \rightarrow \text{大}, x_2 \rightarrow \text{大}, x_3 \rightarrow \text{小}$
- こういう相関関係をどうやって
見つけるか？
 - 主成分分析_(後述)が便利！

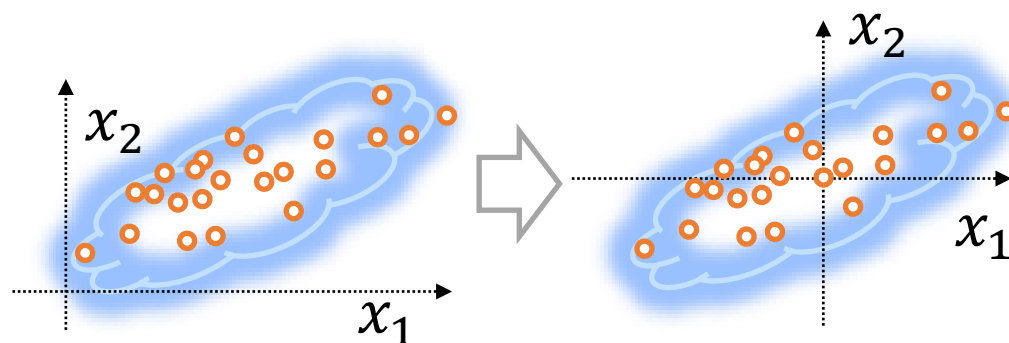


相関係数

相関の程度を測る

相関係数 ρ ～相関の定量化 (1/5)

- 以下では, 簡単のために
 x_1 も x_2 も平均ゼロと仮定
 - =分布をずらしただけ



- この時, 相関係数 ρ は次式!

$$\rho = \frac{(x_1 \cdot x_2) \text{の平均値}}{\sqrt{x_1 \text{の分散} \cdot x_2 \text{の分散}}}$$

分子が大事

分母は正規化, すなわち
 ρ の範囲を $-1 \leq \rho \leq 1$ に
 限定しているだけ

- 次スライドでもう少し詳しく!

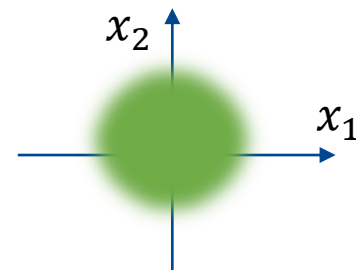
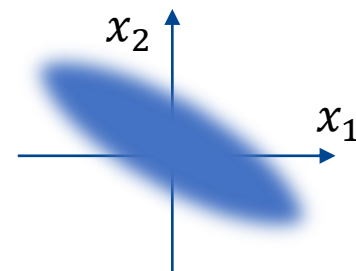
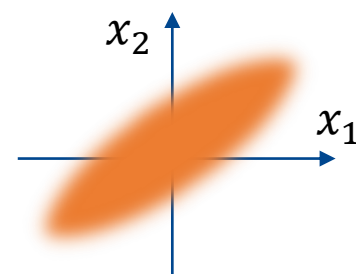


相関係数 ρ ～相関の定量化 (2/5)

- では分子を見てみよう

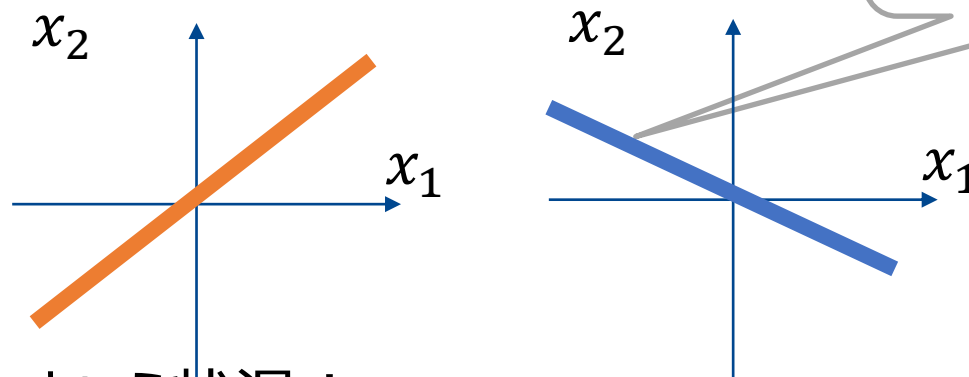
$$\rho = \frac{(x_1 \cdot x_2) \text{の平均値}}{\sqrt{x_1 \text{の分散} \cdot x_2 \text{の分散}}}$$

- x_1 と x_2 が同じ符号(+と-)になりがち
→ $x_1 \cdot x_2$ は正になりがち →その平均(分子)は正
- x_1 と x_2 が逆の符号になりがち
→ $x_1 \cdot x_2$ は負になりがち →その平均(分子)は負
- x_1 と x_2 の符号は同じだったり逆だったり
→ $x_1 \cdot x_2$ も正だったり負だったり →その平均(分子)は0



相関係数 ρ ～相関の定量化 (3/5)

- 例によって「極端な場合」で考えてみよう



- これは $x_2 = ax_1$ という状況！
 - すなわち x_1 が決まれば x_2 の値は ax_1 に一意に決まる状況！
- 相関を計算すると…

$$\begin{aligned}\rho &= \frac{(x_1 \cdot x_2) \text{の平均値}}{\sqrt{x_1 \text{の分散} \cdot x_2 \text{の分散}}} = \frac{(x_1 \cdot ax_1) \text{の平均値}}{\sqrt{x_1 \text{の分散} \cdot ax_1 \text{の分散}}} = \frac{a(x_1 \cdot x_1) \text{の平均値}}{|a|(x_1 \text{の分散})} \\ &= \frac{a(x_1 \cdot x_1) \text{の平均値}}{|a|(x_1 \cdot x_1) \text{の平均値}} = \frac{a}{|a|} = \begin{cases} 1 & a > 0 \\ -1 & a < 0 \end{cases}\end{aligned}$$

相関係数 ρ ～相関の定量化 (4/5)

- 相関 ρ は「 x_1 が定まると x_2 がどれくらい定まるか」の指標でもある

- x_1 と x_2 の相関 ρ が ± 1 (先ほどの状況)

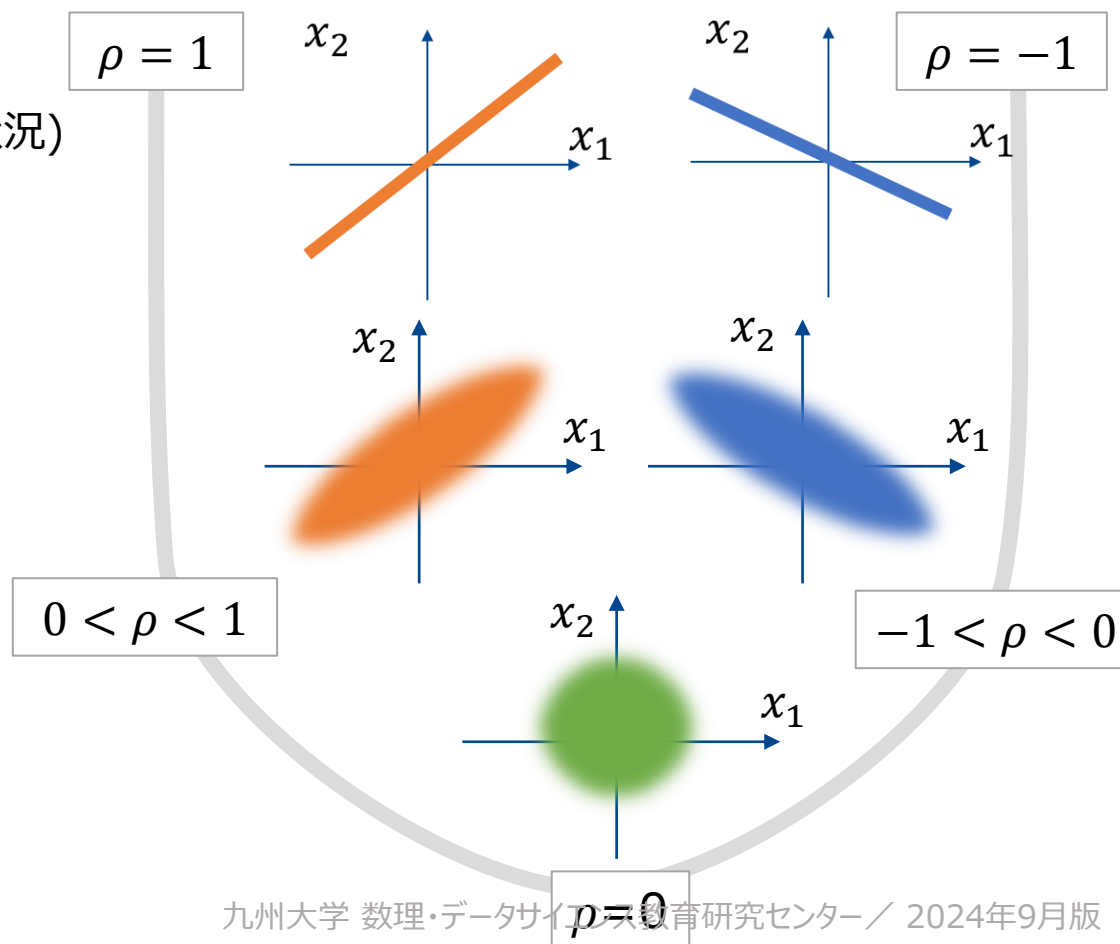
- どちらかが決まれば
他方は一意に定まる

- ± 1 でもなく0でもない

- 緩やかに影響しあっている

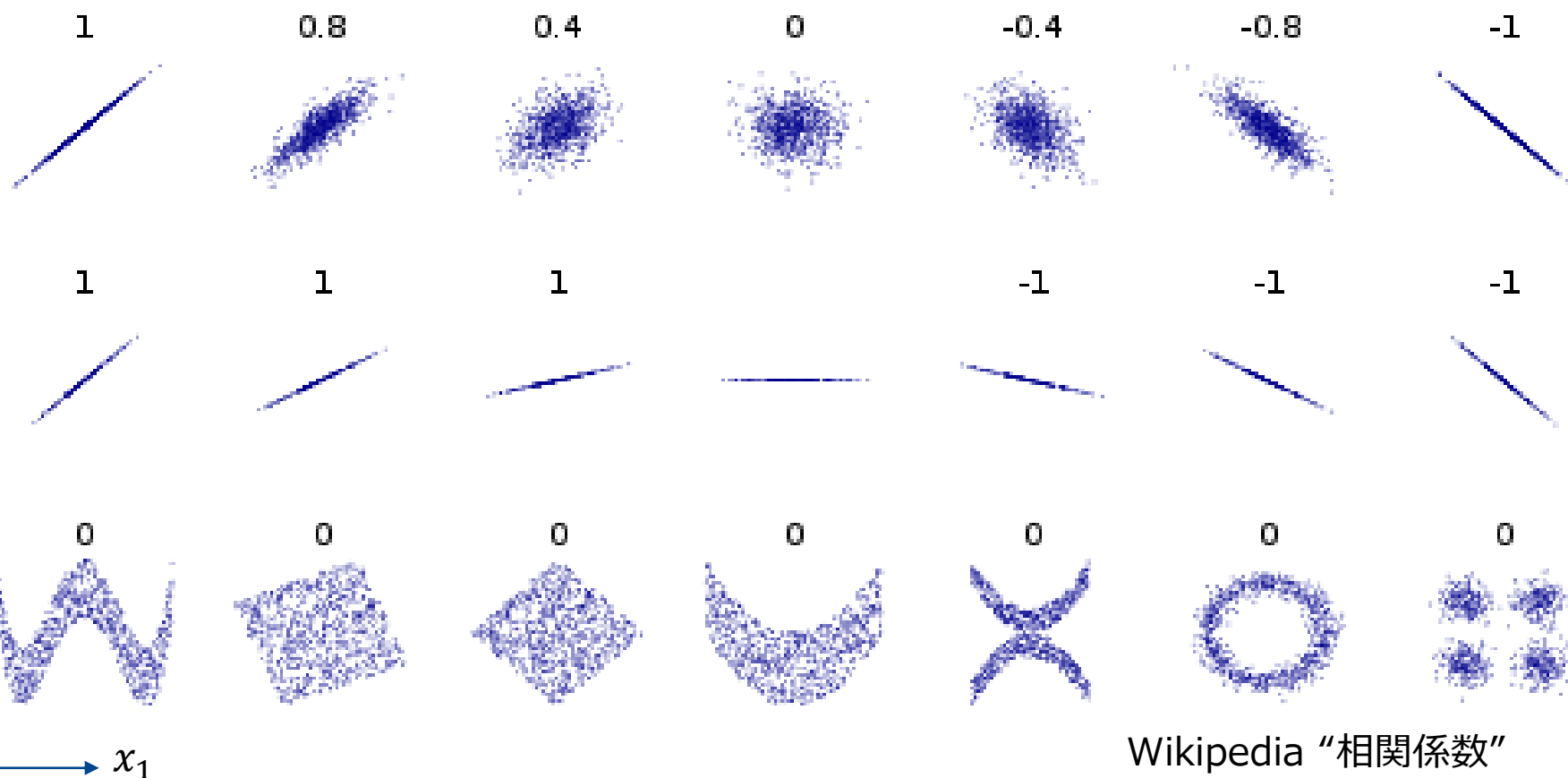
- x_1 と x_2 の相関 ρ が0

- 両者は無相関.
一方の値は他方に影響せず



相関係数 ρ ～相関の定量化 (5/5)

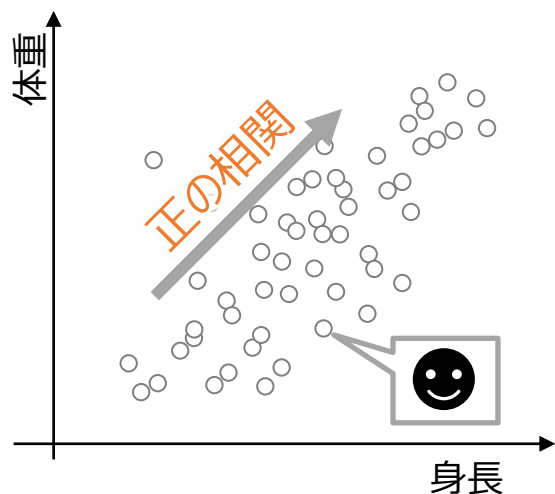
- 以上より, 相関係数 ρ がわかると, 分布の形をある程度想像できる



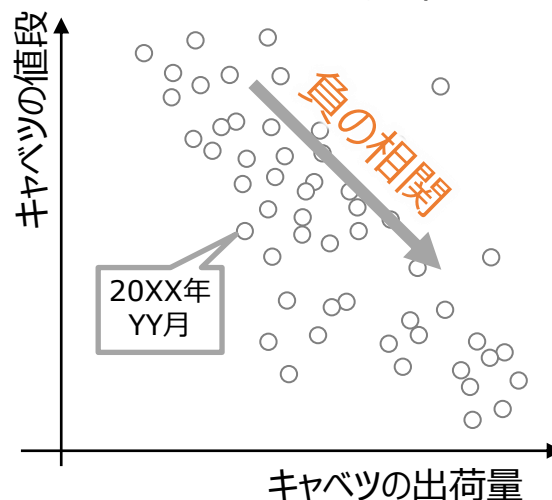
Wikipedia “相関係数”

ここまででわかったこと： 相関には「正の相関・負の相関・無相関」がある

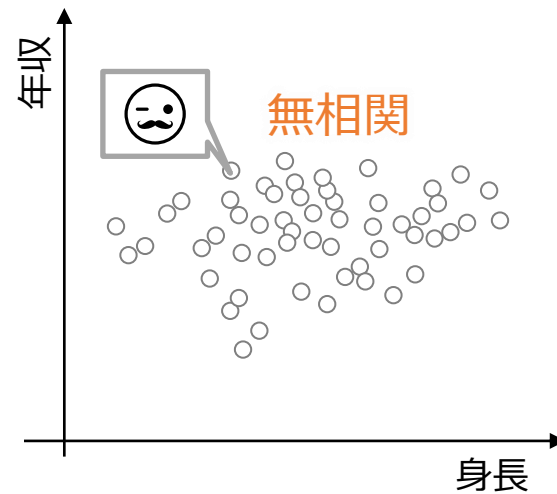
身長が高いほど体重が重い傾向
(正の相関が強い)



出荷量が多いほど価格が下がる傾向
(負の相関が強い)



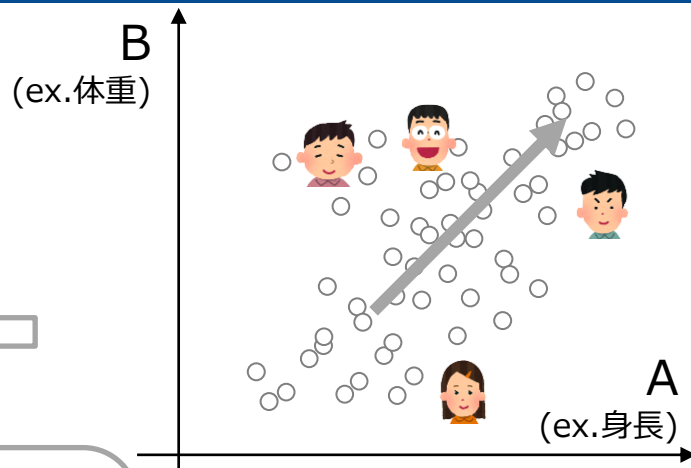
身長と年収の間には特別の傾向は見られない
(相関ゼロ = 最も相関が弱い)



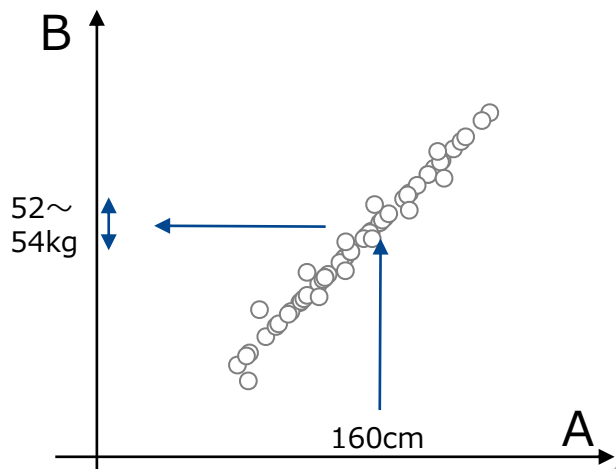
● 考えてみよう

- 「勉強時間」と「テストの点数」は {正の相関, 負の相関, 無相関} ?
- 動画の「長さ」と「データ量」は {正の相関, 負の相関, 無相関} ?
- カレーライスの「分量」と「値段」は {正の相関, 負の相関, 無相関} ? ? ?

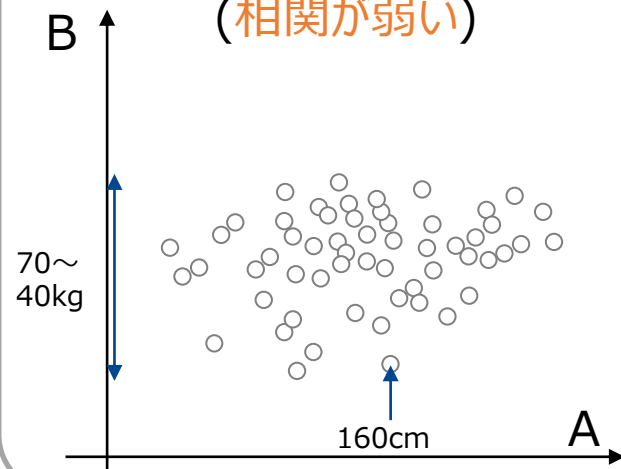
ここまででわかったこと： 相関には「強さ」がある＝傾向には強さがある



傾向が強い
= AをわかるとBも結構わかる
(相関が強い)



傾向が弱い
= Aがわかってても
Bを知るのにはあまり役に立たない
(相関が弱い)

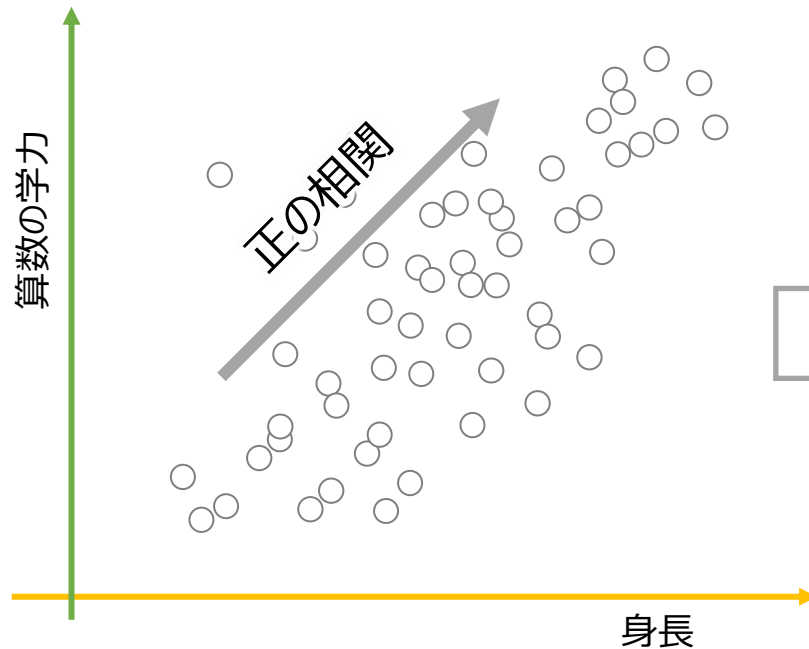


擬似相関

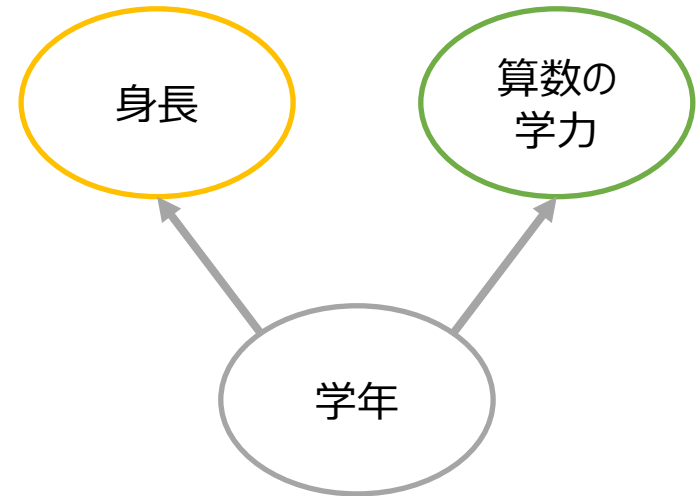
だまされないように！ だまされないように！

擬似相関には気を付けよう！ 背が高いと算数が得意!?

だまされないように
気を付けよう！



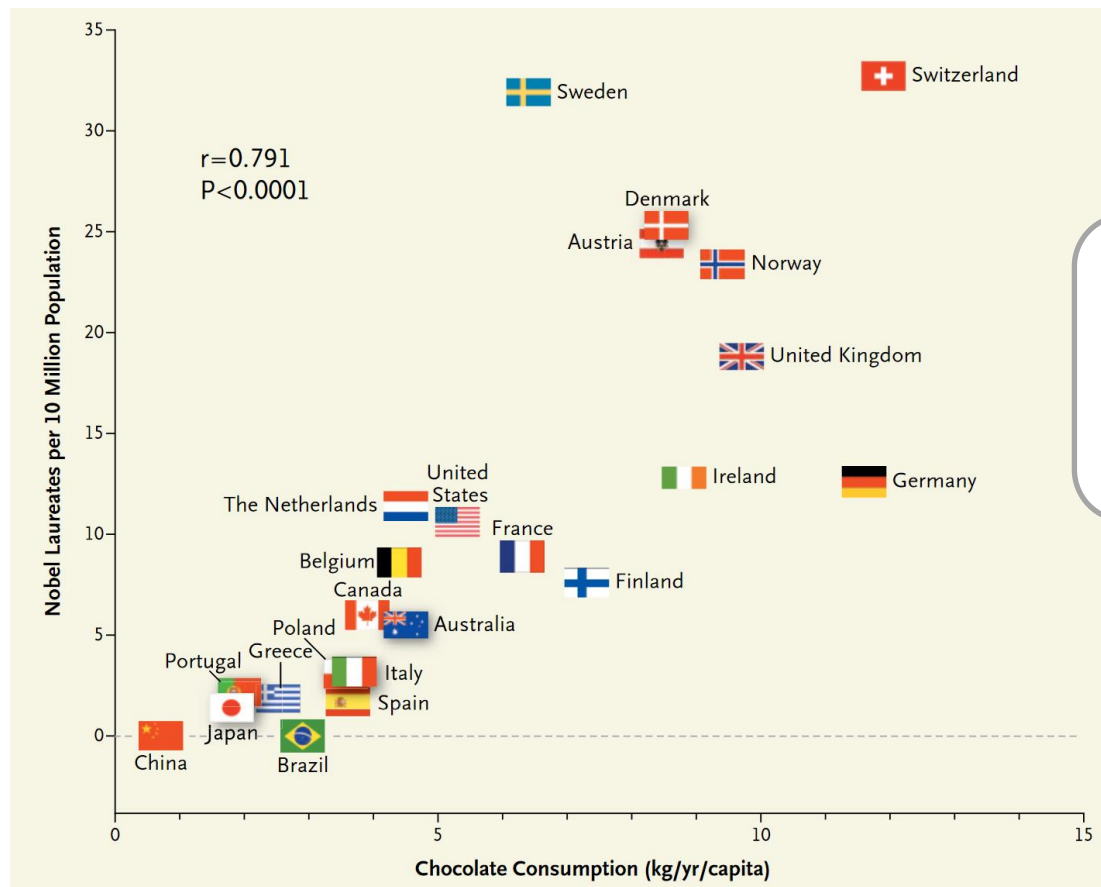
「身長」と「算数の学力」には
正の相関があった！
(背が高いほうが数学得意！)



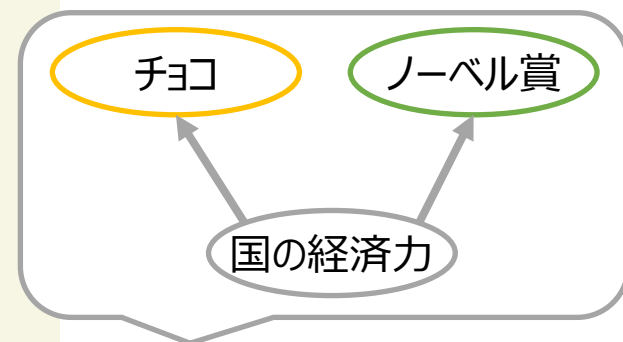
「学年」という要因のために
「**見かけ上**」相関しているだけ

擬似相関には気を付けよう！ チョコを食べる国はノーベル賞が多い!?

人口1千万人あたりのノーベル賞受賞者



チョコレート消費量



「チョコを食べる国はノーベル賞が多い」ことを 本気で証明するのは、結構大変

1. 同じような（年齢，健康状態，食生活，住所，成長過程などが似た）人々を集め，**ランダムに2群**に分ける
2. 第一群はチョコレートを食べる，第二群は食べないという条件以外は，極力同じような状況で過ごしてもらう。



3. 数年後に，2 群の間で，頭脳に差が出るかどうかをテストする
（集めた人々のノーベル賞の数で評価してもよい）

「相関と因果関係は違う」ことにも 気を付けよう！

だまされないように
気を付けよう！



- 因果関係＝「こういう原因だから，こういう結果になった」
- AとBに相関関係があっても，AとBのどちらが原因・結果かは不明
- さらには擬似相関の可能性もあるので要注意



としても…

サプリ

長寿命

サプリを飲んだから長寿命

長寿命

サプリ

長寿命だったから多くのサプリを飲めた

サプリ

長寿命

健康意識

健康意識が高いからサプリも飲むし長寿命

どれが本当かは不明

「サプリ摂取量」と「寿命」には
正の相関があった！

因果推論と効果検証については→付録

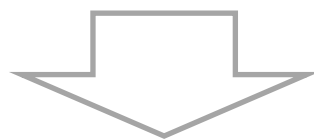
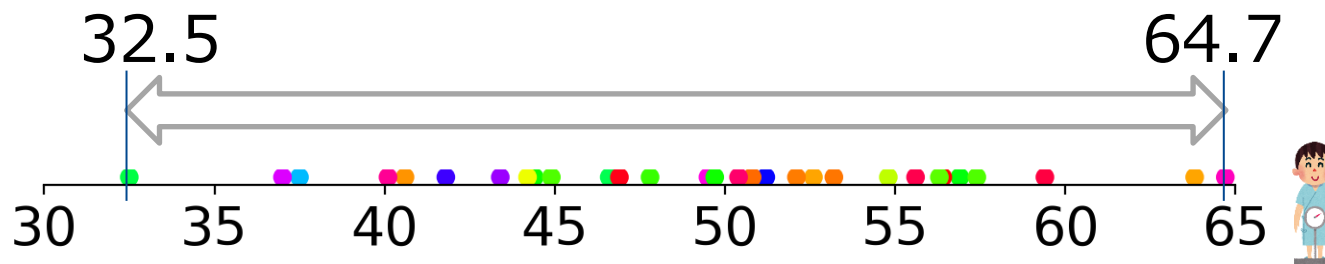
【付録 1】

なぜ「ばらつき」を「分散」で測るのか？
なぜ「分散」では「差の二乗」を使うのか？

そういう疑問を持つことは正しい

なぜ「ばらつき」を「分散」で測るのか？
もっと簡単に「最大－最小」ではどうか？

- 3年10組の「最大－最小」 = $64.7 - 32.5 = \underline{32.2}$

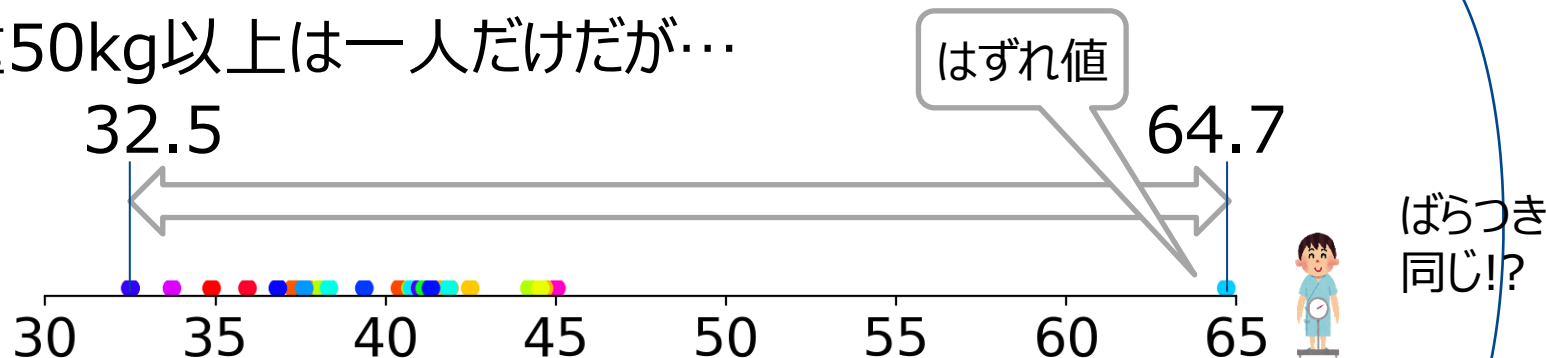


悪くなさそうだけど、
実は「ばらつき」を適切に表現できないケースも

「最大－最小」が「ばらつき」を適切に表現できない例

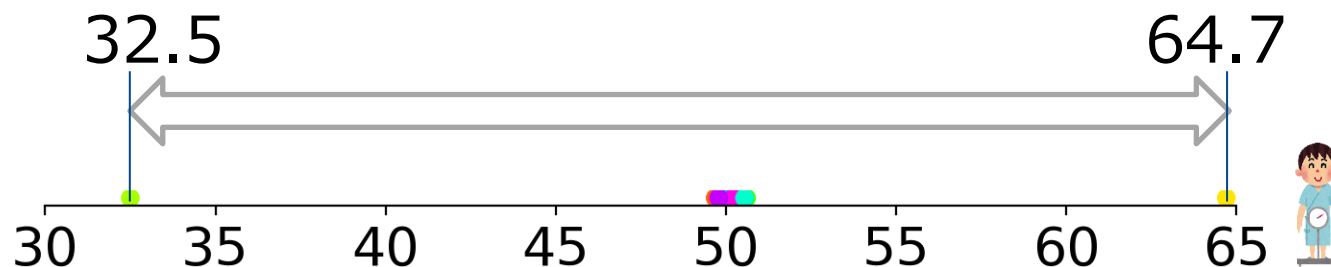
- 1年Q組の「最大－最小」 = $64.7 - 32.5 = \underline{32.2}$

- 体重50kg以上は一人だけだが…



- 3年Z組の「最大－最小」 = $64.7 - 32.5 = \underline{32.2}$

- 30人中28人が50kg付近＝ほとんどばらついてないのに…

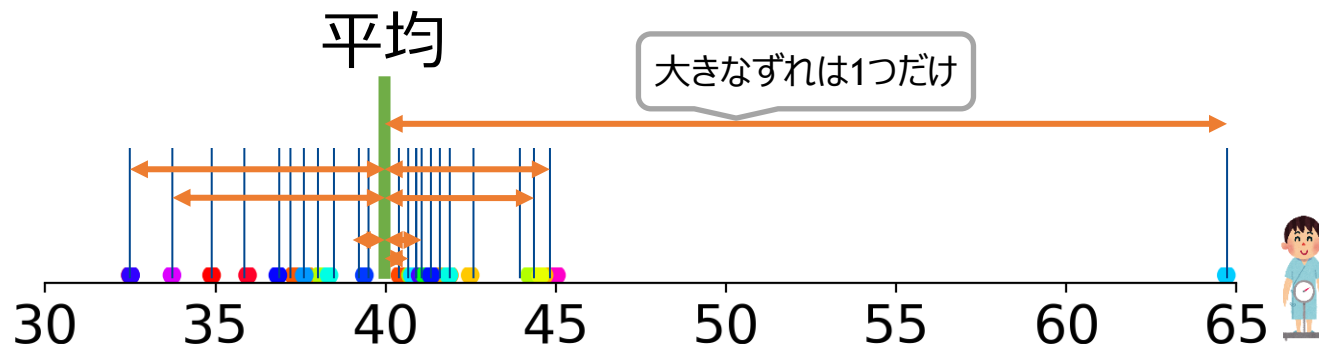


「最大－最小」ではなぜうまくいかないか、 そして解決へのアイデア

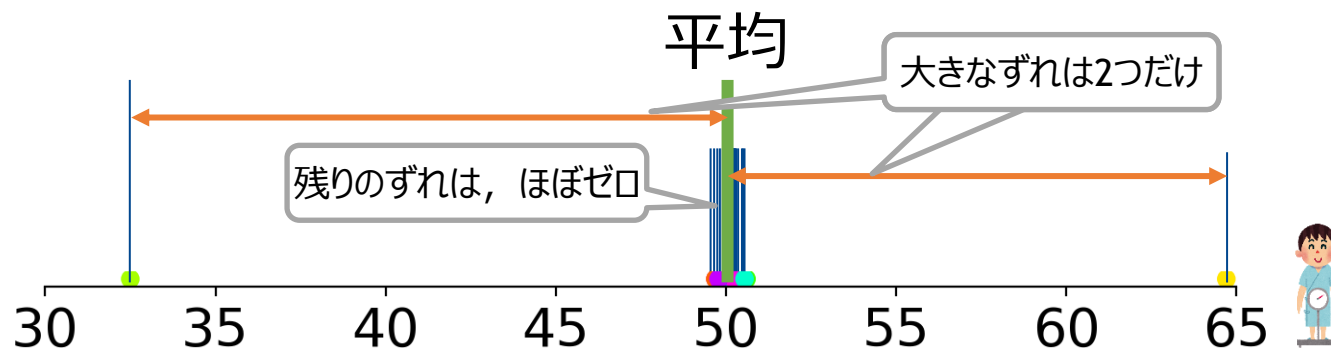
- 「最大－最小」がうまく行かないのは、**たった2つの値(最大値, 最小値)**だけで全データのばらつきを表そうとしているため
 - それら2つが例外的な値であれば、問題が発生
 - さらに悪いことに、最大値・最小値は「例外的」な値になりやすい
- そこで、**全データ**を使って「ばらつき」を計算したい
 - → 「各」データついて「ばらつき」の程度が計算できれば...
 - → そこで「分散」登場！

「最大－最小」でうまく行かなかったケースでも、分散ならば適切に「ばらつき」が求まる

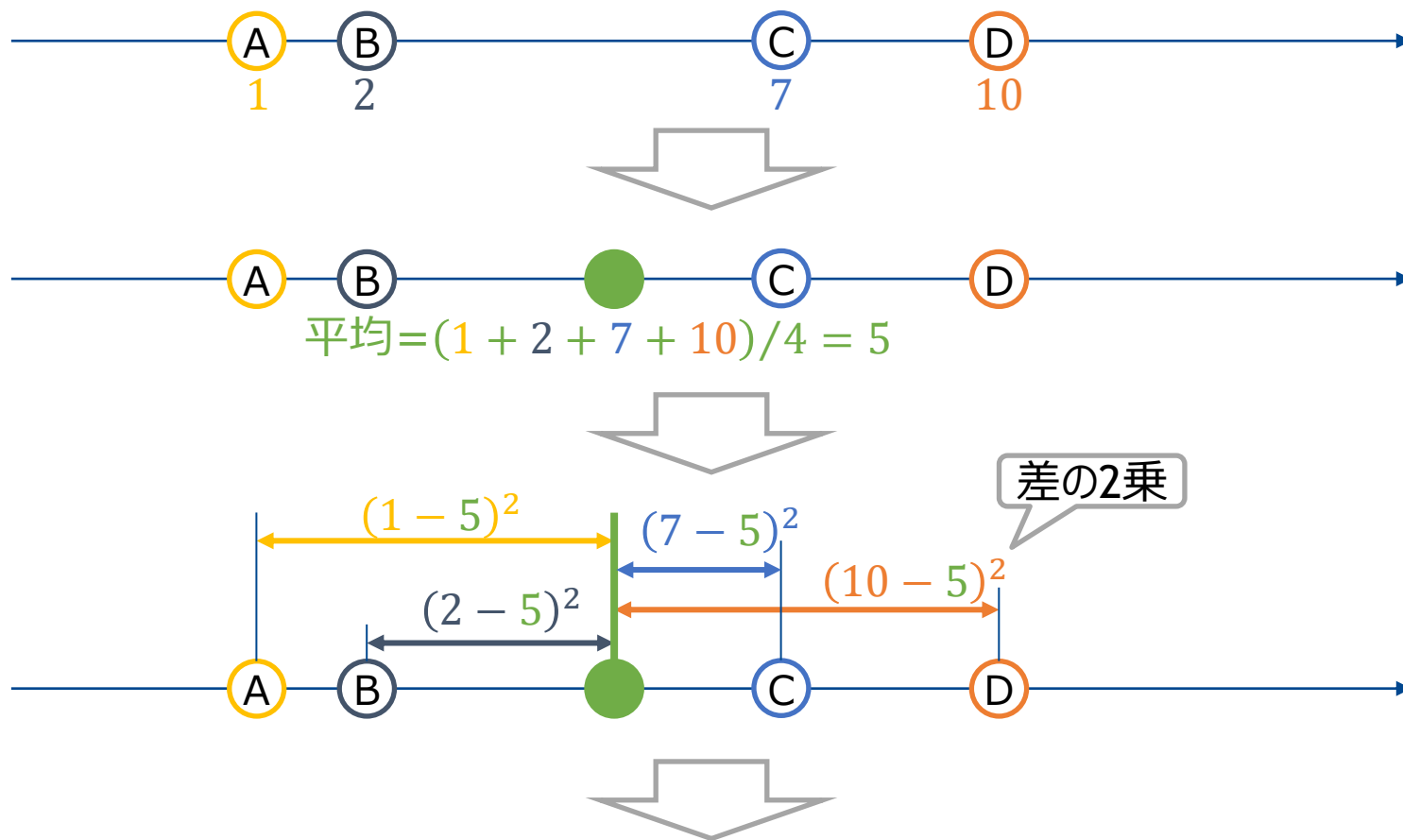
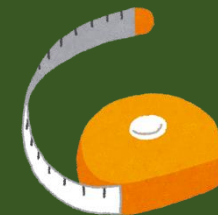
- 1年Q組 → 分散はそう大きくならない



- 3年Z組 → 分散はゼロに近い! → ばらつきは非常に小さい



分散における「ずれ」の測り方： 平均との差の「2乗」で測る

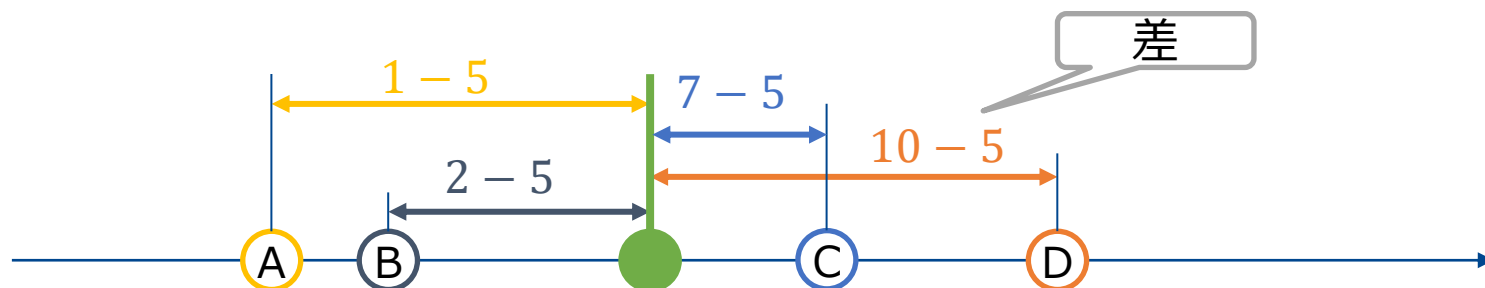


$$\begin{aligned} \text{分散} &= ((1 - 5)^2 + (2 - 5)^2 + (7 - 5)^2 + (10 - 5)^2) / 4 \\ &= (16 + 9 + 4 + 25) / 4 = 13.5 \end{aligned}$$

なぜ「分散」では「差の二乗」を使うのか？ (1/4) 「平均との差」ではだめか？




- 「平均との差」をそのまま平均すると，必ずゼロになる

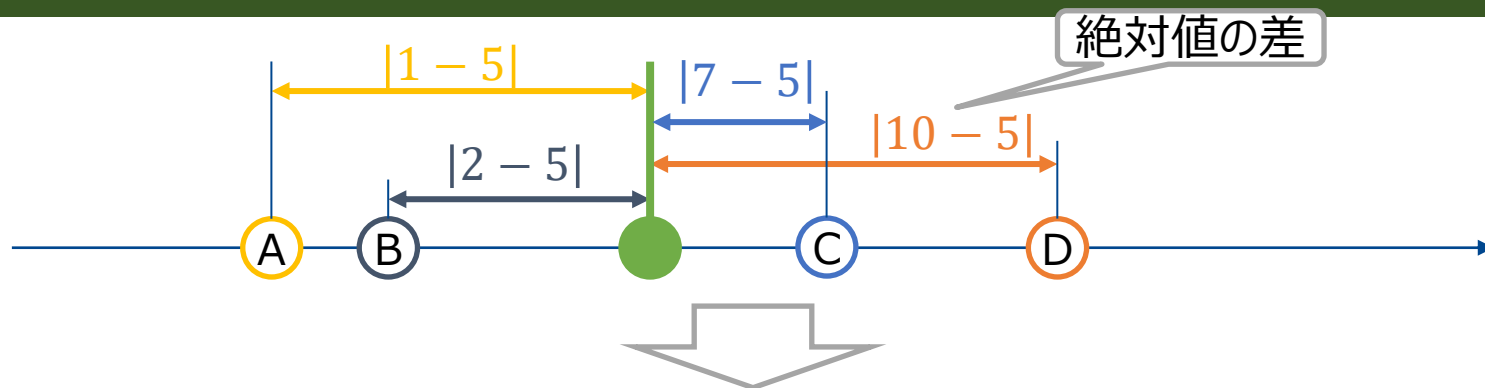


$$\begin{aligned}
 & (1 - 5 + 2 - 5 + 7 - 5 + 10 - 5) / 4 \\
 & = (-4 - 3 + 2 + 5) / 4 \\
 & = 0
 \end{aligned}$$

そりゃそうか…
これじゃばらつきは
測れない



なぜ「分散」では「差の二乗」を使うのか？ (2/4)
 ならば、「差の絶対値」では？ $|1-6| \rightarrow$ 



$$(|1-5| + |2-5| + |7-5| + |10-5|)/4 = (4 + 3 + 2 + 5)/4 = 3.5$$

- 別にこれでもOK (「平均偏差」という名前がついている)

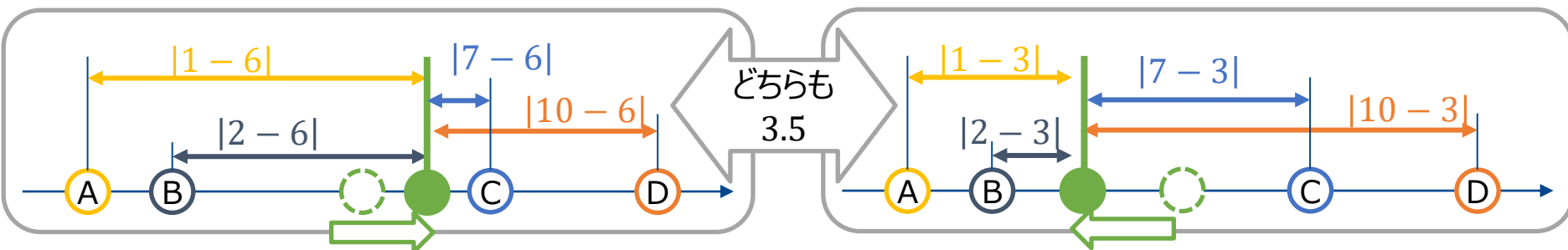
- でも、どうして、平均偏差はあまり使われないのだろう



- 次スライドにその理由がいくつか...

なぜ「分散」では「差の二乗」を使うのか？ (3/4) 平均偏差はなぜあまり使われない？

- 「平均より大」のデータと「小」のデータが同数の場合、「平均が移動しても変わらない」ので、やや解釈がむずかしい



- 絶対値だと扱いづらい（微分不可能だし）
- 正規分布の分散値を最尤推定すると、二乗の式が出てくる

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$$

- さらにもう一つ...

これはちょっと(今は)
よくわからない



なぜ「分散」では「差の二乗」を使うのか？ (4/4)

分散と平均の美しい関係が導かれる

- 分散（バラつき）の基準点，**本当はどこがいいのか？**

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2 \quad \xrightarrow{\text{平均にしてたけど}} \quad f(a) = \frac{1}{N} \sum_{i=1}^N (x_i - a)^2 \quad \text{一旦忘れてみる}$$

- バラつきが小さくなるような基準点を求めるとすると...

$$\frac{d f(a)}{d a} = \frac{d}{d a} \left[\frac{1}{N} \sum_{i=1}^N (x_i - a)^2 \right] = -\frac{2}{N} \sum_{i=1}^N (x_i - a) = 0$$

$$\Leftrightarrow \sum_{i=1}^N (x_i - a) = 0 \Leftrightarrow N a = \sum_{i=1}^N x_i \Leftrightarrow a = \frac{1}{N} \sum_{i=1}^N x_i = \text{平均!}$$

- というわけで「『平均』が分散の基準点としてふさわしい」ことも保証！
 - 差の二乗で測ったからからこそ導かれる，平均と分散の美しい関係
 - 平均偏差 $\frac{1}{N} \sum_{i=1}^N |x_i - \bar{x}|$ にこのような関係はない

【付録 2】 偏差値

状況が違ってても、どれぐらいずれているかを比べたい

A君は〇〇模試で80点, B君が△△模試で80点,
どちらも自分がどれぐらいスゴイのか知りたい

- A君の受けた〇〇模試の平均点=60点



- B君の受けた△△模試の平均点=70点



- そうすると平均より20点上だった, A君のほうがスゴそう

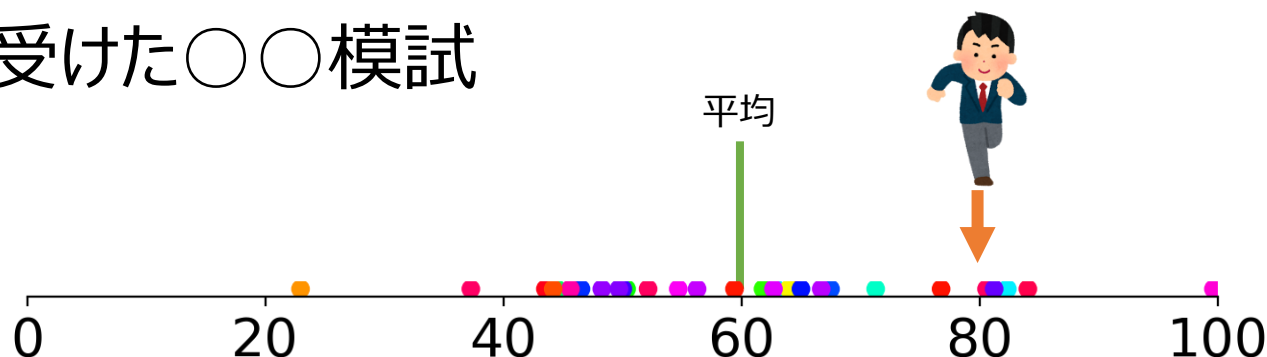


- でも, もし, △△模試のほぼ全員が70点ぐらいだったら, その状況で80点とったB君のほうがすごい?

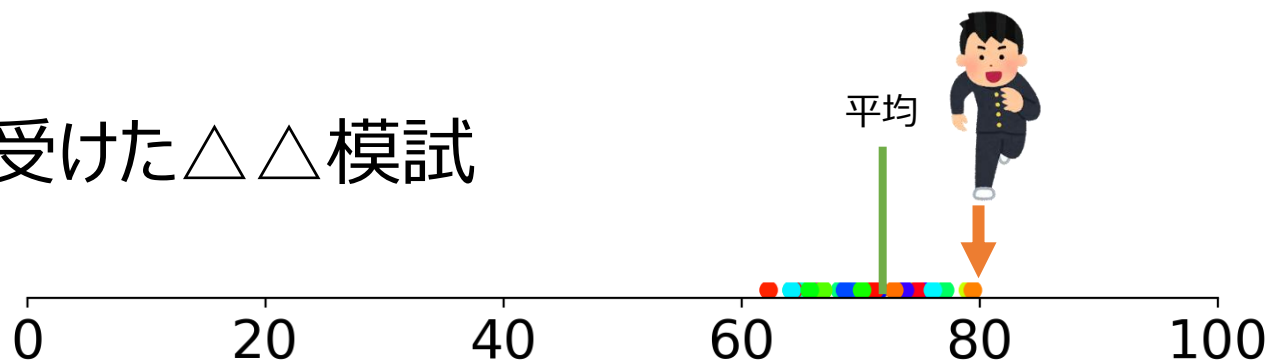


点数の分布を見てみると...

- A君の受けた〇〇模試



- B君の受けた△△模試



- 確かに，B君のほうがスゴイかも...

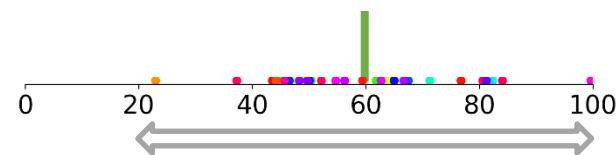
ではどうやって二人のすごさを比較するか？

- 平均との差だけではダメっぽい

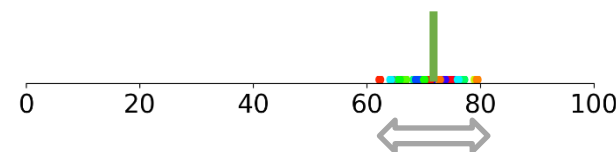
- そこで、「ばらつき」具合を使う！



- ○○模試のように、「ばらつき」が大きい
→ 平均との差を小さめに評価
= 平均と結構違っても「あまり変わらない」



- △△模試のように、「ばらつき」が小さい
→ 平均との差を大きめに評価
= 平均と少しでも違えば「すごく違う」



実際にはどうする？

「平均からのずれ」を標準偏差で割ればOK

$$\frac{\text{データの値} - \text{平均}}{\text{標準偏差}}$$

平均からのずれ

ばらつき具合

- ○○模試の標準偏差=20 (ばらつき大)
- △△模試の標準偏差=5 (ばらつき小)
- A君の値 = $(80-60)/20 = 1$
- B君の値 = $(80-70)/5 = 2 \rightarrow$ B君のほうが、より平均から離れている!

偏差値



$$\text{偏差値} = 10 \times \frac{\text{データの値} - \text{平均}}{\text{標準偏差}} + 50$$

- ○○模試の標準偏差=20 (ばらつき大)
- △△模試の標準偏差=5 (ばらつき小)

- A君の値 = $10 \times (80 - 60) / 20 + 50 = 60$
- B君の値 = $10 \times (80 - 70) / 5 + 50 = 70$

B君のほうが
偏差値高い



偏差値の性質



$$\text{偏差値} = 10 \times \frac{\text{データの値} - \text{平均}}{\text{標準偏差}} + 50$$

- 平均点と同じなら偏差値50
- 標準偏差が小さい(ばらつきが小さい)ほど, データの少しの変化が大きく影響する

因果推論と効果検証の基礎

ダイエットでやせたのは、本当にそのサプリのおかげ？



因果推論と効果検証

- 因果推論

- 「ある原因がある結果を引き起こしているのかどうか」を明らかにする

- 効果検証

- 「意図的に与えた原因」の結果への影響（効果）を明らかにする
- 例：「サプリを飲んだら」（意図的に与えた原因）→「体重が減った」（結果）



- 基本的考え方

- 原因の有無で結果がどう変わるかをチェック

- 効果検証の方法

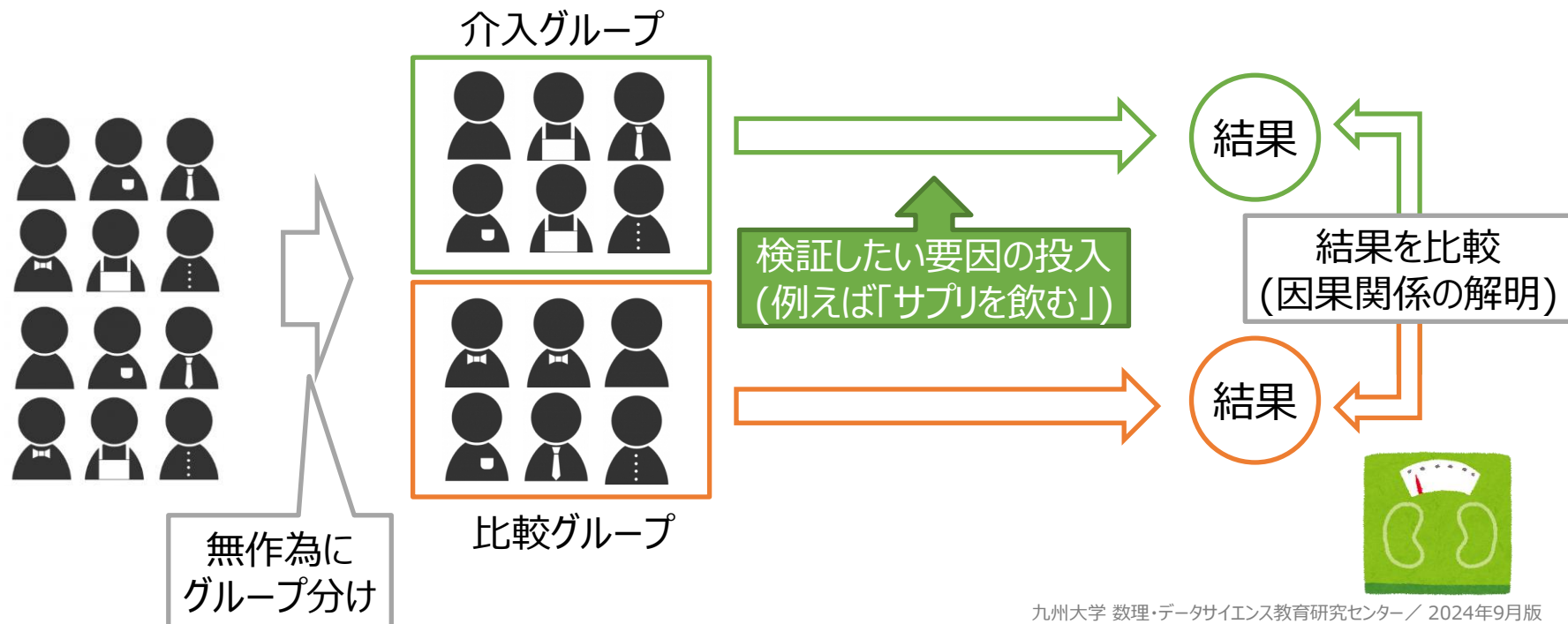
- 原因の有無を積極的に作る方法 → ランダム化比較試験とA/Bテスト
- 勝手にできた「原因の有無」を利用する方法 → 自然実験

「原因の有無を積極的に作る」効果検証法(1)

ランダム化比較試験

Randomized Controlled Trial (RCT)

- 検証したい要因以外は公平になるように，対象の母集団を無作為にグループに分け，その検証したい要因の影響や効果を明らかにするための比較方法
- 具体的には「介入グループ」と「比較グループ」の2種類に分け，その試験的要素の影響を測定



ランダム化比較試験の例

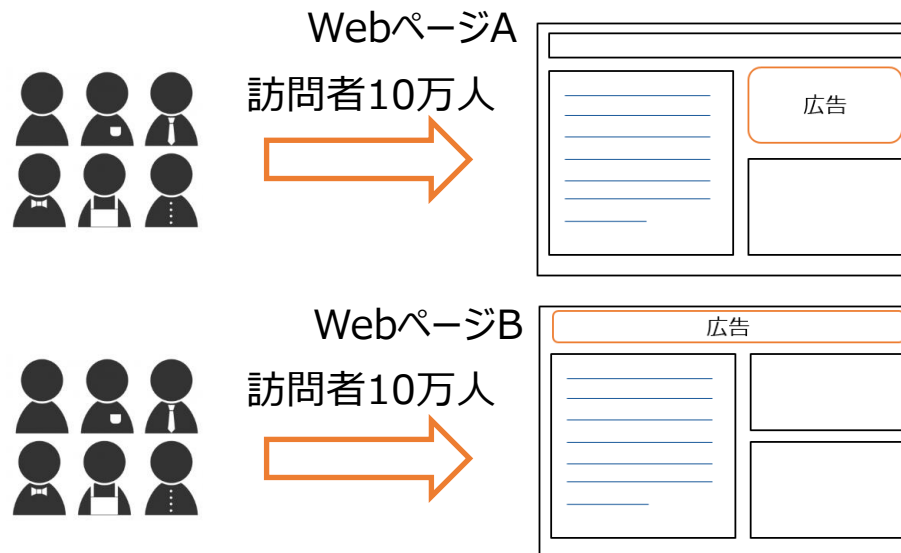
- 電力価格を上げると本当に節電につながるのか？
 - 参加者: 北九州市内の一般参加世帯
 - 介入群: 電力の需給が特にひっ迫する数時間の間, 節電を促すための価格上昇を経験
 - 明らかにしたい因果関係: 電力消費量に差が出れば, 電力価格の上昇が電力消費量に影響を及ぼす
- 実験結果から
 - 電力価格の上昇は節電を促すという因果関係が分かった
 - 料金を上げるほど, 価格の上昇に応じて節電が進む

参考: 依田高典 田中誠 伊藤公一朗, 「スマートグリッド・エコノミクス フィールド実験・行動経済学・ビッグデータが拓くエビデンス政策」
<http://www.iwafunelab.iis.u-tokyo.ac.jp/crest/20161121/20161121-4.pdf>

「原因の有無を積極的に作る」効果検証法(2)

A/Bテスト

- 2通り(以上)のパターンを用意し, どちらがより効果が高い成果が出るのかを検証する方法
 - インターネットのマーケティング分野で主に使われる
 - ランダム化比較試験の考え方を基礎にしている
 - オバマ氏も大統領選挙でより多くの支援者を獲得するために活用した手法
- 例) どのように広告を掲載すると (原因) , クリック率が上がる? (結果)



どちらの広告配置がより多くの人にクリックされたか(広告商品の売り上げに貢献したか)を調査

ランダム化比較試験（A/Bテスト）の問題点と、 解決策としての「自然実験」

- 実験に必要となる費用や労力などが膨大
- 各グループに十分な数の調査対象が必要
- 状況によっては、ランダム化比較試験を実施できない
 - 医療費の自己負担額を変化させると、医療サービスの利用頻度にどのような影響があるか
 - 所得税を低くすると、その国(地域)に移住する人は増えるのか

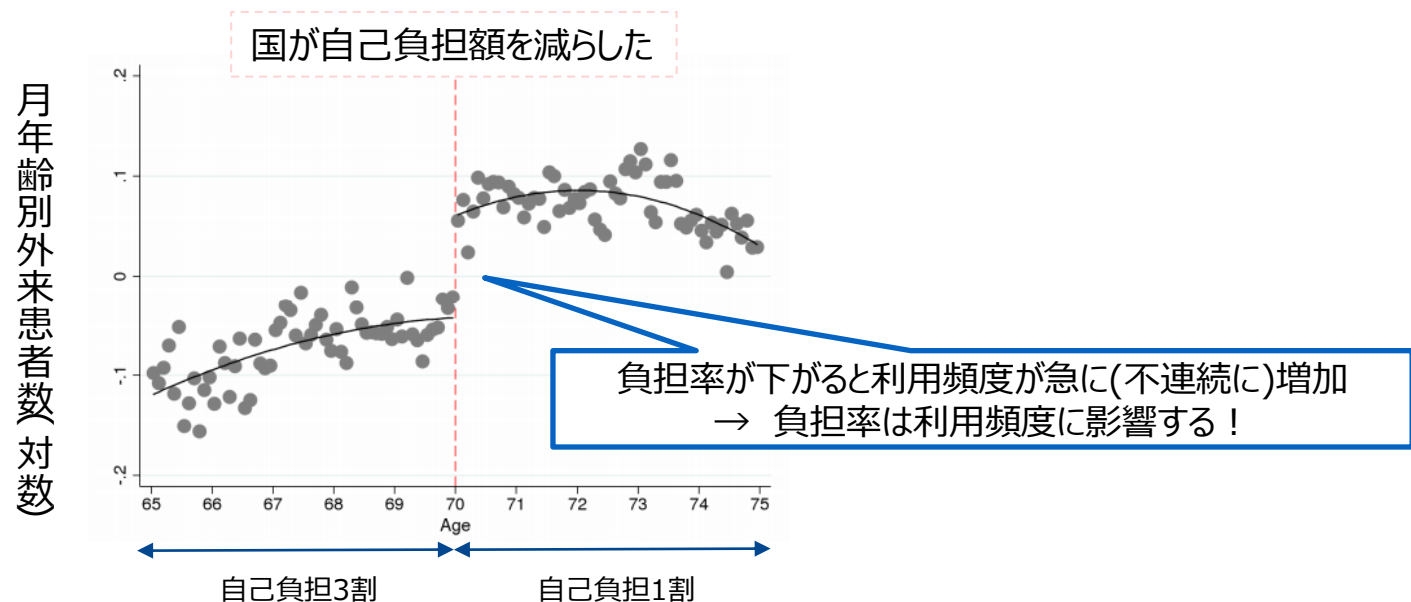
こんな実験は
倫理的にも社会的にも
難しい

- 自然実験を利用
 - 自然実験＝「自然に（＝勝手に）、比較実験と同じような状況ができた」
 - その状況を「うまく」見つけて使って、効果検証する

「自然実験」による効果検証の例： RDデザイン

Regression Discontinuity (RD) design

- 世の中に(調査とは無関係に)発生した「不連続」を用いた効果検証
 - 例：「国が医療費の自己負担率を下げた」ことを利用して,
 - 医療費負担率（原因）と医療サービス利用頻度（結果）の関係を調査



Hitoshi Shigeoka, 2014. "The Effect of Patient Cost Sharing on Utilization, Health, and Risk Protection," American Economic Review, American Economic Association, vol. 104(7), pages 2152-84